UNIVERSITA' DEGLI STUDI DI PADOVA

LAUREA SPECIALISTICA IN SCIENZE STATISTICHE ECONOMICHE, FINANZIARIE E AZIENDALI

ANALISI ROBUSTA DELLA SOPRAVVIVENZA NELLO STUDIO DEL MESOTELIOMA MALIGNO

RELATORE: **Prof.ssa LAURA VENTURA**
Dipartimento di Scienze Statistiche

LAUREANDA: **ELISABETTA MATTIOLO**

Anno Accademico 2011/2012

Ai miei genitori

La condotta dei genitori
è la guida dei figli.

Proverbio

INDICE

INTRODUZIONE

L'utilizzo di dati reali per l'elaborazione di una tesi, a mio avviso, è di fondamentale importanza per poter verificare e sperimentare praticamente quanto appreso nella mia carriera universitaria di studente alla Facoltà di Scienze Statistiche, dapprima con la Laurea Triennale in Scienze Statistiche e Tecnologie Informatiche ed ora con la Laurea Specialistica.

Quindi, la possibilità di elaborare dati "veri", provenienti dall'archivio della Sezione di Anatomia Patologica del Dipartimento di Scienze Medico-Diagnostiche e Terapie Speciali dell'Università degli Studi di Padova (cfr. Cappellesso, 2009), riguardanti il mesotelioma maligno, una neoplasia che sta rapidamente aumentando in tutto il mondo a causa della diffusa esposizione all'amianto, ha subito destato il mio interesse.

In Italia, come pure in Nuova Zelanda e in Francia, si osserva un tasso grezzo d'incidenza del mesotelioma maligno compreso tra 1.1 e 2.0 casi per 100.000. Il più elevato tasso grezzo d'incidenza proviene dall'Australia, Belgio e Gran Bretagna con circa 3 casi per 100.000 persone/anno. Mentre in altri paesi, quali Canada, USA, Giappone, paesi dell'Europa Centrale e Orientale è invece inferiore a 1.1 casi (Bianchi e Bianchi 2007).

E così, oltre ad approfondire le mie conoscenze assai limitate su questo particolare tumore in espansione in Italia, e che vede interessata anche la Regione Veneto, ho avuto l'opportunità di applicare le varie nozioni e tecniche di elaborazione dati acquisite in questi anni universitari.

Lo schema della tesi è il seguente.

Nel Capitolo 1 viene illustrato, con opportuni riferimenti medici, il tumore "mesotelioma maligno" per poter comprendere meglio la successiva analisi preliminare dei dati e le relative elaborazioni effettuate utilizzando il software *R - Version 2.11.1*. Infatti, nel Capitolo 2 viene proposta l'analisi esplorativa delle variabili osservate. Vengono riportati grafici e relativi commenti per una dettagliata analisi del campione di pazienti a cui è stata diagnosticato questo tumore, derivante dall'esposizione all'amianto.

Quindi si prosegue, nel Capitolo 3, con l'analisi della variabile dipendente "mesi di sopravvivenza" utilizzando modelli per l'analisi della sopravvivenza quali, lo stimatore di *Kaplan-Meier* e il Modello di *Cox* a rischi proporzionali.

Infine, si propone nel Capitolo 4 l'analisi dei dati con l'uso di un modello robusto per dati di sopravvivenza (Heritier *et al.*, 2009), cioè di metodi meno sensibili alla presenza di eventuali dati "anomali" o osservazioni influenti.

Nelle Conclusioni vengono riepilogati i vari metodi di studio applicati per i dati in esame e quindi riassunti in breve i risultati ottenuti dalle varie elaborazioni.

CAPITOLO 1

IL MESOTELIOMA MALIGNO

Il mesotelioma maligno (MM) è un grave forma di cancro correlata all'esposizione alle fibre aerodisperse dell'amianto (asbesto), derivante dalle cellule di rivestimento (mesotelio) delle cavità sierose quali la pleura, il peritoneo, il pericardio e la tunica vaginale.

Si presenta macroscopicamente come un ispessimento della pleura, generalmente diffuso, più raramente nodulare.

In Italia, dal 2003, presso l'Istituto Superiore per la Prevenzione e la Sicurezza sul Lavoro (ISPESL) è attivo il Registro Nazionale dei Mesoteliomi (ReNaM), ad articolazione regionale, al fine di stimare l'incidenza di MM in Italia, definire le modalità di esposizione, l'impatto e la diffusione della patologia nella popolazione e di identificare sorgenti ancora ignote di contaminazione ambientale da amianto.

Proprio dal rapporto ReNaM (Terracini, 2006) si apprende che questa tipologia di tumore può essere causata dall'ambiente lavorativo per gli operatori impegnati nella produzione e nell'utilizzo industriale di amianto e derivati, o può essere paraoccupazionale, per l'uso dei relativi manufatti o raramente per esposizione in locazioni geologiche a polveri di origine naturale, non di cava.

Esiste una relazione inversa tra l'intensità dell'esposizione all'asbesto e la lunghezza del periodo di latenza; tuttavia, generalmente la neoplasia si sviluppa dopo un lungo periodo d'esposizione. Inoltre, il rischio di sviluppare il MM aumenta con la durata dell'esposizione.

E' ormai riconosciuto che il periodo di tempo che intercorre tra la prima esposizione all'asbesto e la diagnosi di MM, è di 20 - 40 anni.

L'incidenza di questa neoplasia appare in crescita in tutto il mondo con circa 2.2 casi per milione di abitanti (dati epidemiologici dal sito *www.gime.it/clinica03.htm*). Essendo fortemente correlata all'uso industriale dell'amianto, attualmente vietato ed in fase di eliminazione in alcuni paesi, ed essendo la patologia ad alta latenza temporale, si prevede un picco di casi intorno al 2020, ed una successiva decrescita.

In Italia, l'estrazione, l'importazione, l'esportazione, la commercializzazione e la produzione di amianto e di prodotti contenenti amianto è stata vietata del 1992 (in riferimento alla Legge n° 257/92).

Essendo, però, il nostro paese uno dei maggiori produttori ed utilizzatori di amianto fino alla fine degli anni '80, attualmente ne deve sopportare le conseguenze dei livelli di esposizione causati dall'uso intenso del materiale dal secondo dopoguerra nei settori della produzione industriale di manufatti in cemento-amianto, di manufatti tessili contenenti amianto, della cantieristica navale, della riparazione e demolizione di rotabili ferroviari e dell'edilizia.

Secondo l'Associazione Italiana per la Ricerca sul Cancro, il MM nel nostro paese rappresenta lo 0.4% di tutti i tumori diagnosticati nell'uomo e lo 0.2% di quelli diagnosticati nelle donne. Ciò equivale a dire che si verificano 3.4 casi di MM ogni 100.000 uomini e 1.1 ogni 100.000 donne.

1.1 Classificazione

A seconda del distretto corporeo nel quale ha origine, il MM è classificato in:

- **Mesotelioma pleurico:** si genera nella cavità toracica e rappresenta la tipologia più diffusa (circa 3 casi su 4);
- **Mesotelioma peritoneale:** nasce nell'addome e rappresenta la quasi totalità dei mesoteliomi rimasti, escludendo quelli pleurici;
- **Mesotelioma pericardico:** nasce nella cavità attorno al cuore ed è estremamente raro;
- **Mesotelioma della tunica vaginale:** nasce dalla membrana che riveste i testicoli ed è molto raro.

Se invece si prende in considerazione il tipo di cellula maligna presente nel tumore, secondo la classificazione dell'Organizzazione Mondiale della Sanità, si distinguono tre tipi di mesotelioma (vedi Figura 1.1):

- **Epitelioide**: nella maggior parte le cellule hanno citoplasma che, all'interno, presenta delle granulazioni di colore rosato**,** con nuclei relativamente regolari. E' il più comune (60-70% dei casi) e quello che tende ad avere una migliore prognosi;

- **Sarcomatoide** o fibroso: formato da cellule fusate arrangiate in fasci o distribuite in maniera disordinata. Rappresenta dal 10 al 20% dei mesoteliomi;
- **Misto** o bifasico: nel 30% dei casi i mesoteliomi comprendono sia aspetti epitelioidi che sarcomatoidi. Può essere presente qualsiasi combinazione dei *patterns* suddetti. Ciascuna componente deve rappresentare almeno il 10% della neoplasia per legittimare tale termine.

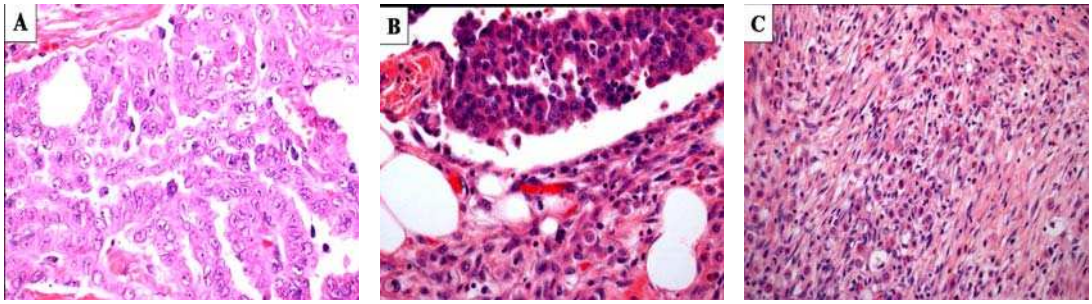

Figura 1.1: MM di tipo epitelioide (A), bifasico (B) e sarcomatoso (C). Ingrandimenti originali 200x e 400x.

Nel caso del MM, i due sistemi di stradiazione, processo attraverso il quale si valuta l'estensione locale e la diffusione a distanza di un tumore, più comunemente utilizzati (dal sito *www.gime.it/clinica09.htm*) sono il *Butchart Staging System* ed il *TNM Staging System*. Entrambi individuano quattro stadi e si riferiscono alla forma pleurica, in relazione alla sua più frequente osservazione.

Il *Butchart Staging System*, risalente all'anno 1979, si basa unicamente sull'estensione della massa tumorale primitiva, mentre la più recente classificazione stilata nell'anno 2009 dall'*American Joint Commitee on Cancer* (AJCC), ossia la "*TNM Staging System*", risulta più accurata in quanto considera tre principali variabili: le dimensioni della massa primitiva del tumore (T), l'eventuale coinvolgimento dei linfonodi (N) e la presenza di metastasi ad organi distanti (M). L'unione di questi diversi parametri consente la suddivisione in stadi che descrivono pertanto l'esatta estensione della neoplasia. Il primo stadio configura un MM localizzato, mentre i tre successivi stadi si riferiscono ad una malattia avanzata:

- Stadio I: Tumore limitato alla pleura destra o sinistra. Può interessare il polmone, il pericardio o il diaframma omolaterali. Non c'è coinvolgimento linfonodale.

- Stadio II: Tumore esteso ai linfonodi peribronchiali e/o ilari omolaterali. Si può essere già diffuso al polmone, al diaframma o al pericardio omolaterali.
- Stadio III: Tumore che invade i muscoli della parete toracica, le costole, il cuore, l'esofago o altri organi nel torace omolaterali, con o senza interessamento dei linfonodi sub-carenali e/o mediastinici omolaterali.
- Stadio IV: Tumore con interessamento dei linfonodi intratoracici eterolaterali (dal lato opposto alla massa primitiva) e/o tumore che coinvolge, per estensione diretta: - pleura e/o polmone eterolaterali - peritoneo e/o organi addominali - strutture del collo.

La stadiazione costituisce uno dei fattori che maggiormente orientano la scelta della strategia terapeutica da adottare e che permettono di stabilire la prognosi.

1.2 Sintomi e cure

I sintomi del mesotelioma sono inizialmente molto poco specifici e spesso vengono ignorati o interpretati come segni di altre malattie più comuni e meno gravi. L'assenza di specificità, accentuata soprattutto nei primi stadi, rende conto del lungo intervallo di tempo intercorrente tra l'esordio della sintomatologia ed il momento della diagnosi.
La metà circa dei pazienti scoprono la reale natura della loro malattia dopo più di sei mesi dalla comparsa dei sintomi.
I segni precoci del mesotelioma pleurico possono includere dolore nella parte bassa della schiena o a un lato del torace, fiato corto, tosse, febbre, stanchezza, perdita di peso, difficoltà a deglutire, debolezza muscolare. Dolore addominale, perdita di peso, nausea e vomito sono invece sintomi più comuni in caso di mesotelioma peritoneale.
Lo strumento più efficace per confermare il sospetto di mesotelioma è la *biopsia*.
In alcuni casi con un ago lungo e sottile vengono prelevati campioni di liquido presenti nel torace (toracentesi), nell'addome (paracentesi) o nella cavità attorno al cuore (pericardiocentesi) e si verifica al microscopio la presenza di cellule tumorali. In altri casi, invece, è necessario prelevare piccole porzioni di tessuto mesoteliale con un ago sottile inserito sottopelle o con l'inserimento di una sonda dotata di videocamera attraverso un piccolo taglio nella pelle: in questo modo il medico può vedere le aree sospette e prelevare i campioni che vengono poi analizzati al microscopio. Per distinguere con certezza il mesotelioma da altri tipi di tumore, i campioni prelevati con la biopsia possono essere sottoposti ad analisi immunoistochimiche (per vedere le

proteine presenti sulla superficie della cellula) o genetiche (per individuare l'espressione di geni tipica del mesotelioma).

L'esame Tomografia Assiale Computerizzata (TAC) permette di determinare la presenza del tumore, la sua posizione esatta e la sua eventuale diffusione ad altri organi, aiutando anche il chirurgo a definire il tipo di trattamento più adatto.

Determinare lo stadio del tumore è infatti essenziale per valutare la possibilità di *intervenire chirurgicamente*: un tumore "resecabile", cioè che può essere asportato con la chirurgia, ha infatti più probabilità di essere curato rispetto a uno non operabile.

Comunemente nello stadio I, e in limitati casi di stadi II e III, è prevista la *pleurectomia* con la decorticazione della stessa oppure l'asportazione sia della pleura che della parte di polmone coinvolti *(pleuropneumectomia)* o, nel caso di localizzazioni peritoneali, l'asportazione del peritoneo colpito *(peritonectomia).*

Tali possibilità d'intervento non dipendono solo dalle dimensioni del tumore, ma anche dal sottotipo, dalla sua posizione e dalle condizioni del paziente.

Per le persone che, per diversi motivi, non possono essere sottoposte a intervento chirurgico si opta per la *radioterapia* che, talvolta, può avere solo scopo palliativo e aumentare la sopravvivenza di solamente qualche mese. In alcuni casi, questo tipo di terapia può essere usata anche dopo la chirurgia (radioterapia adiuvante) per distruggere i piccoli gruppi di cellule tumorali non visibili e quindi non asportabili nel corso dell'operazione.

Un altro tipo di cura possibile è la *chemioterapia*, ovvero la somministrazione di un farmaco con un'iniezione intravenosa che lo porta in tutto l'organismo, o direttamente nella cavità toracica (per via intrapleurica) oppure addominale (per via intraperitoneale). Ciò può contribuire a rallentare la progressione della malattia anche se difficilmente riesce a curarla in modo definitivo.

Attualmente non esiste una cura specifica e risolutiva al MM. Si tratta infatti di un tumore raro a cui viene associata un'alta mortalità e quindi non è semplice per i medici confrontare l'efficacia dei diversi trattamenti o avere l'esperienza necessaria per compiere la scelta giusta.

Come per la maggior parte dei tumori, anche per il MM, minore è lo stadio, maggiori sono le probabilità di buona riuscita del trattamento, anche se spesso la diagnosi di questo tumore arriva quando la malattia ha già superato gli stadi iniziali e risulta difficile da trattare.

Ciò è confermato dagli studi del ReNaM (Terracini, 2006): il periodo di sopravvivenza di un paziente a cui è stato diagnosticato il MM è tra i 4 e 12 mesi per la sede pleurica e varia tra i 5 e 35 mesi per quella peritoneale.

1.3 EMT: Transizione epiteliale - mesenchimale

La transizione epiteliale - mesenchimale (EMT) è un processo in cui, in seguito ad uno stimolo cronico, le cellule epiteliali con polarità basale - apicale perdono il loro fenotipo e acquisiscono le caratteristiche delle cellule mesenchimali non polarizzate e in grado di migrare (Yang e Weinberg, 2008).

L'EMT è un punto chiave dello sviluppo embrionale e, recentemente, è stato assodato che tale processo è la base dello sviluppo e dell'evoluzione dei tumori in generale.

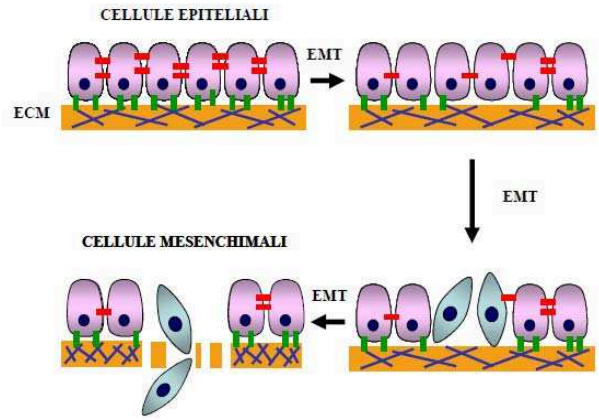

Figura 1.2: Transazione epiteliale – mesenchimale.

Come si può vedere dalla Figura 1.2, in condizioni di normalità, le cellule epiteliali sono saldamente legate le une alle altre e alla membrana basale tramite giunzioni aderenti e giunzioni occludenti (*tight junctions*).

Le prime sono presenti a livello basolaterale e sono caratterizzate dalla presenza di proteine transmembranarie, appartenti alla famiglia delle caderine. Le *tight junctions*, localizzate a livello basale, riconoscono una struttura analoga ma sono costituite da proteine transmembranarie specifiche (occludine e claudine) e legate a proteine del citoscheletro. Le molecole che compongono le giunzioni cellulari non svolgono esclusivamente una funzione di adesione meccanica, ma sono anche responsabili del mantenimento dell'integrità strutturale e della polarità cellulare.

A seguito dell'alterazione a carico delle giunzioni cellulari, le cellule perdono di coesione sia tra loro, sia rispetto alla membrana basale, perdendo la polarità funzionale delle diverse superfici cellulari e acquisendo mobilità.

Il processo EMT permette infatti ad una cellula epiteliale di una matrice extracellulare ECM (cellule altamente polarizzate e connesse le une alle altre da giunzioni cellulari) di distaccarsi dal sito del tumore primario acquistando la capacità di migrare convertendosi in cellule di fenotipo mesenchimale, caratterizzate da legami deboli e da strutture irregolari.

Durante la EMT, le cellule epiteliali perdono le giunzioni intercellulari, con conseguente distacco dalle altre cellule circostanti e acquisiscono caratteristiche mesenchimali necessarie per migrare lontano dal sito del tumore primario.

Diverse serie di EMT e il loro opposto, vale a dire la transizione epiteliale-mesenchimale (MET), sono necessarie per la differenziazione finale dei vari tipi cellulari e per l'acquisizione della complessa struttura tridimensionale degli organi interni dell'embrione.

Queste serie sequenziali sono state suddivise in EMT primaria, secondaria e terziaria (si veda Thiery *et al.*, 2009).

La primaria include la formazione, a seguito del processo di gastrulazione, dell'ultimo dei tre foglietti germinativi, il mesoderma, e la delaminazione delle cellule della cresta neurale dal tubo neurale dorsale, da cui derivano le strutture craniofacciali, la maggior parte del sistema nervoso periferico, alcune cellule endocrine e i melanociti (Thiery, 2002).

Le cellule mesodermiche precoci si suddividono in assiali, parassiali, intermedie e della placca laterale e condensano in strutture epiteliali transitorie, attraverso una transizione mesenchimale-epiteliale: i somiti, i precursori del sistema urogenitale, la somatopleura e la splancnopleura. Queste strutture secondarie vanno incontro ad un'altra EMT che porta alla formazione di cellule mesenchimali con un potenziale di differenziazione più ristretto, da cui originano tipi cellulari specifici: i dermatomeri, i miotomi, gli sclerotomi, il palato e il tratto riproduttivo (Thiery *et al.*, 2009).

L'EMT terziaria, infine, avviene a livello del canale atrioventricolare e coinvolge le cellule endoteliali che così invadono la gelatina cardiaca e formano il cuscino endocardico, precursore delle valvole cardiache.

L'EMT si dimostra sempre più una tappa chiave nella progressione della neoplasia verso la metastasi, ma il suo ruolo non si limita all'invasione e alla migrazione. Infatti,

l'EMT è coinvolta anche nella resistenza all'apoptosi, all'anoikis e alla senescenza, nell'immuno-tolleranza e immuno-soppressione, nella farmacoresistenza ed infine conferisce proprietà staminali. Generalmente si ritiene che le metastasi derivino da cellule di una neoplasia avanzata, diventate in grado di invadere e disseminare.
Studi recenti tuttavia indicano che il processo di EMT e la conseguente disseminazione di cellule neoplastiche possono avvenire in maniera continua dall'inizio dello sviluppo del tumore primitivo e non solamente in stadi avanzati, per lo meno in alcuni tipi di neoplasie.

1.3.1 I marcatori dell'EMT

Sono stati identificati dei marcatori molecolari per valutare se una cellula epiteliale è andata o meno incontro al processo EMT. Il principale marcatore è la perdita della *E-caderina*, evento associato alla distruzione delle giunzioni cellula-cellula.
Le *caderine* sono una famiglia di glicoproteine transmembrana, localizzate a livello delle giunzioni aderenti, che mediano il contatto intercellulare attraverso legami omotipici con caderine della stessa specie dipendenti dal calcio. L'*E-caderina* (caderina epiteliale) è necessaria per la formazione di giunzioni aderenti forti e stabili e quindi per il mantenimento del fenotipo epiteliale e della normale architettura tessutale dell'adulto. Una riduzione d'espressione dell'*E-caderina*, com'è stato evidenziato in varie neoplasie, svolge un ruolo cruciale nella perdita della differenziazione cellulare e nella disseminazione (Hirohashi, 1998); si può dire che l'*E-caderina* appare come la custode del fenotipo epiteliale. La *N-caderina* (caderina neurale) normalmente si trova espressa solo nelle cellule del sistema nervoso, ma viene prodotta in alcune cellule di carcinoma che hanno perso l'espressione *dell'E-caderina* e, in questo contesto cellulare, è associata ad un aumentato potenziale invasivo (Nieman *et al.*, 1999).
Nell'EMT si assiste spesso a questo avvicendamento di caderine tra *l'E-caderina* e la *N-caderina*. La funzione delle caderine dipende dalla loro associazione con il citoscheletro d'actina ed è mediata per mezzo dell'interazione tra la regione C-terminale delle caderine stesse e le proteine citoplasmatiche chiamate catenine. La stabilità dell'associazione tra le caderine e il citoscheletro d'actina è regolata dalla fosforilazione e defosforilazione della *β-catenina*. Questa è anche coinvolta nella regolazione dell'espressione genica come mediatore della via di trasduzione del segnale Wnt, il

quale controlla la sorte cellulare, e nell'EMT, in cui la *β-catenina* può essere presente all'interno del nucleo.

Vi è inoltre la superfamiglia di fattori di trascrizione *SNAIL, ZEB e TWIST* che si attivano durante il processo EMT. In particolare *SNAIL 1* e *2* sono in grado di reprimere la trascrizione dell'*E-caderina*, di aumentare l'espressione e l'attività delle metalloproteinasi della matrice (*MMP*), di mediare l'EMT e l'invasione in linee cellulari di neoplasie umane.

La famiglia di fattori di trascrizione nominata *ZEB*, di nostro particolare interesse, è composta da due membri, *ZEB1* e *ZEB2*, codificati da due geni indipendenti. Questi fattori sono caratterizzati dalla presenza di due insiemi di 3 o 4 *zinc finger*, una regione centrale ricca di serine e proline e una N-terminale diversa per ciascuna proteina, a ciascun'estremità e di un *homeodomain* al centro (Nieto, 2002). I membri della famiglia *ZEB* interagiscono con il DNA attraverso il legame simultaneo dei due domini *zinc finger* ai siti di legame ad alta affinità composti da precisi elementi ripetuti bipartiti (CACCT e CACCTG), come quelli che si trovano nel promotore del gene della *E-caderina* (CDH1), ma il cui orientamento e distanza possono variare considerevolmente in bersagli diversi. *ZEB2* reprime la trascrizione dell'*E-caderina* legandosi al suo promotore e aumenta l'espressione delle *MMP* mediando così l'EMT e l'invasione in alcune neoplasie umane.

Le *MMP* sono una famiglia di zinco endopeptidasi, la cui espressione e attività sono aumentate nella EMT, che sono responsabili della degradazione delle componenti dell'EMC svolgendo un ruolo cruciale nell'invasione neoplastica e nella metastasi.

SNAI1, *SNAI2*, *ZEB1*, *ZEB2* e *TWIST1* piuttosto che essere ridotti a semplici repressori della trascrizione dell'*E-caderina*, andrebbero visti come regolatori del fenotipo epiteliale, dell'adesione e del movimento cellulare.

Il *citoscheletro,* che è formato dai microfilamenti d'actina, dai filamenti intermedi e dai microtubuli, svolge un ruolo indispensabile nella migrazione mesenchimale, essendo responsabile della resistenza agli stress meccanici e della deformabilità della cellula. Nell'EMT il citoscheletro d'actina si caratterizza per la scomparsa della disposizione circolare dei fasci, tipicamente epiteliale, e per la distribuzione lineare degli stessi, come nelle cellule mesenchimali.

Questa profonda riorganizzazione del citoscheletro, essenziale per acquisire la motilità, è accompagnata dal cambiamento delle componenti dei microfilamenti: si passa infatti da una struttura epiteliale composta da β-actina e γ-actina a una mesenchimale

composta da α-actina. I filamenti intermedi sono particolarmente collegati alla funzione fisiologica della cellula, mostrano spiccate differenze molecolari e sono espresse in programmi tessuto-specifici.

Per esempio le citocheratine (filamenti intermedi di tipo I e II) definiscono i tessuti epiteliali, mentre la vimentina (filamento intermedio di tipo III) definisce un'origine mesenchimale: durante l'EMT la vimentina sostituisce le citocheratine.

La proteina *S100A4* (chiamata anche *Fibroblast Specific Protein-1, FSP-1*) è un membro della famiglia delle proteine leganti il calcio associate al citoscheletro normalmente espresse nelle cellule fibrose (mesenchimali), ma non nelle cellule epiteliali. A parte un ruolo nella trasduzione del segnale del calcio, *S100A4* è considerata un marcatore dell'EMT, in quanto modula la motilità cellulare attraverso l'orientamento e la localizzazione delle protrusioni cellulari, è implicata nella crescita e nella differenziazione cellulare ed è coinvolta nell'invasione aumentando l'espressione e l'attività delle *MMP* (Garrett *et al.*, 2006).

Per determinare l'espressione di questi marcatori della transizione epiteliale - mesenchimale possono esser utilizzate le seguenti metodiche:

- l'immunoistochimica *IHC*, tecnica qualitativa e semi-quantitativa che serve per evidenziare, in una sezione di tessuto, determinati anticorpi;
- la PCR quantitativa *Real Time qRT-PCR,* in grado di seguire la reazione di polimerizzazione, finché questa avviene, grazie all'utilizzo di molecole fluorescenti che permettono di determinare quantitativamente il livello in cui una specifica molecola di DNA è presente in un campione.

1.4 Obiettivi dello studio

Dall'archivio della Sezione di Anatomia Patologica del Dipartimento di Scienze Medico-Diagnostiche e Terapie Speciali, dell'Università degli Studi di Padova, sono stati recuperati ed esaminati 76 blocchetti istologici con diagnosi di MM, provenienti da pazienti sottoposti a biopsia o a resezione chirurgica, tra il 2002 e il 2010.

Tutte le diagnosi di MM sono state confermate da dati clinici, morfologici e immunoistochimici secondo i criteri dell'Organizzazione Mondiale della Sanità.

Per poter effettuare un'accurata elaborazione dei dati tramite il *software* statistico *R*, le varie informazioni pervenute dall'osservazione di questi casi clinici a cui è stato diagnosticato il MM, sono state ordinate nel *dataset* in Tabella A.1 (Appendice A).

Obiettivo principale dello studio di questi dati è analizzare la sopravvivenza dei pazienti dalla diagnosi del MM, determinando se la variabile dipendente tempo di sopravvivenza è influenzata dalle caratteristiche, quali il genere, l'età, il tipo di patologia e la sottoclassificazione del tumore riscontrato.

Quindi dopo un'analisi esplorativa dei dati e un'illustrazione generale delle variabili del dataset (Capitolo 2), si procede con l'analisi della variabile d'interesse (Capitolo 3), che rappresenta il tempo di sopravvivenza calcolato in mesi dalla diagnosi del tumore all'eventuale decesso. Viene successivamente proposto (Capitolo 4), in alternativa all'analisi tradizionale basata sul modello di Cox, un approccio robusto per l'analisi della sopravvivenza.

Nel *dataset* sono presenti diverse variabili. La variabile `eta` rappresenta l'età del paziente al momento della diagnosi del tumore. La variabile quantitativa `survival`, indica i mesi di sopravvivenza calcolati dalla diagnosi del tumore al paziente fino al termine dello studio, fissato per il 31 dicembre 2010.

E' quindi doveroso precisare che, come avviene frequentemente in ambito clinico, il *dataset* presenta dati censurati relativi a quei pazienti per i quali il tempo all'evento è superiore alla data di temine dell'osservazione; essi sono definiti censurati a destra e vengono quindi definiti "usciti vivi" dallo studio.

Le variabili di natura qualitativa sono:

- `site`, localizzazione del distretto corporeo in cui si è originato il MM, peritoneale o pleurico;
- `subtype`, sottoclassificazione istologica del MM nei tre livelli: epitelioide, sarcomatoide e misto.

Vi sono quindi due variabili *dummy* (o dicotomiche):

- `sesso`, ovvero se il paziente è di genere maschile (M) o femminile (F);
- `status`, che indica lo stato del paziente (vivo o deceduto) al 31/12/10, termine del periodo di osservazione.

Inoltre, tra le variabili qualitative, ci sono gli 11 marcatori della transizione epiteliale - mesenchimale (EMT):

- `ecad`, E-caderina o caderina epiteliale;
- `ncad`, N-caderina o caderina neurale;
- `bcad`, β-catenina o subunità del complesso proteico delle catenine,
- `mmp2`, zinco endopeptidasi MMP-2;

- mmp9, zinco endopeptidasi MMP-9;
- cyto5.6, filamento intermedio epiteliale citocheratine 5/6;
- vim, filamento intermedio mesenchimale vimentina;
- SMA, actina epiteliale αSMA;
- zeb1, repressore trascrizionale con *zinc finger* ZEB 1;
- zeb2, repressore trascrizionale con *zinc finger* ZEB 2;
- S100A, proteina S100A4 legante il calcio fibroblastica,

La colorazione IHC di questi marcatori è stata valutata con una scala da 0 a 3 con: 0 = 0-5% di media di cellule neoplastiche positive in 5 campi a 400x, 1 = 6-33%, 2 = 34-66%, 3 = 67-100%:

CAPITOLO 2

I DATI

Obiettivo di questo capitolo è presentare una prima analisi generale di tutte le variabili osservate (§ 1.4) per i 76 pazienti sottoposti a biopsia o a resezione chirurgica, tra il 2002 e il 2010. Ai metodi grafici, quali boxplot e istogrammi, si affiancano opportuni test. Nel seguito, in particolare, verranno considerati i test non parametrici basati sui ranghi (si veda ad es. Piccolo, 1998). Si precisa che si lavora con un livello di significatività del 5%.

2.1 Caratteristiche dei pazienti

Il *dataset* utilizzato è riportato in Tabella A.1 dell'Appendice A. In seguito, nelle Tabelle 2.1 e 2.2, sono riportati i riassunti delle variabili, ottenuti con *R*.

VARIABILI QUANTITATIVE							
VAR	**MIN.**	**1° Quart.**	**Mediana**	**Media**	**3° Quart.**	**MAX.**	**sd**
survival	0.00	6.00	11.00	19.72	25.00	82.00	22.12
eta	47.00	64.00	70.00	69.28	76.00	83.00	8.82

Tabella 2.1: Riepilogo delle variabili quantitative.

VARIABILI QUALITATIVE					
Marcatori	**MIN.**	**1° Quart.**	**Mediana**	**3° Quart.**	**MAX.**
ecad	0.000	0.000	1.000	2.000	3.000
ncad	0.000	1.000	2.000	2.000	3.000
bcad	0.000	1.000	2.000	2.000	3.000
mmp2	0.000	0.000	0.000	1.000	3.000
mmp9	0.000	0.000	1.000	2.000	3.000
cyto5.6	0.000	0.000	0.000	1.000	3.000
vim	0.000	1.000	2.000	3.000	3.000
SMA	0.000	0.000	0.000	1.000	3.000
zeb1	0.000	1.000	2.000	3.000	2.000
zeb2	0.000	0.000	0.000	1.000	3.000
S100A	0.000	1.000	1.000	2.000	3.000

VARIABILE	**categorie**	**N°oss.**	**%**
sesso	F	25	32.90
	M	51	67.10
site	Peritoneal	18	23.68
	Pleural	58	76.32
subtype	Epithelioid	46	60.53
	Mixed	14	18.42
	Sarcomatous	16	21.05
status	Deceduto	59	77.63
	vivo	17	22.37

Tabella 2.2: Riepilogo delle variabili qualitative.

Il campione è composto da 51 maschi (67.1%) e 25 femmine (32.9%). L'età, considerata al momento della diagnosi, è compresa tra i 47 e gli 83 anni, con una media di 69 anni (sd = 8.82). Dal boxplot dell'età (Figura 2.1), si può vedere che la fascia d'età considerata critica è tra i 64 e i 76 anni (1° e 3° quartile). Si contano 7 casi (9.2%) di ultra-ottantenni e una decina di "*under-60*" (13.2%).

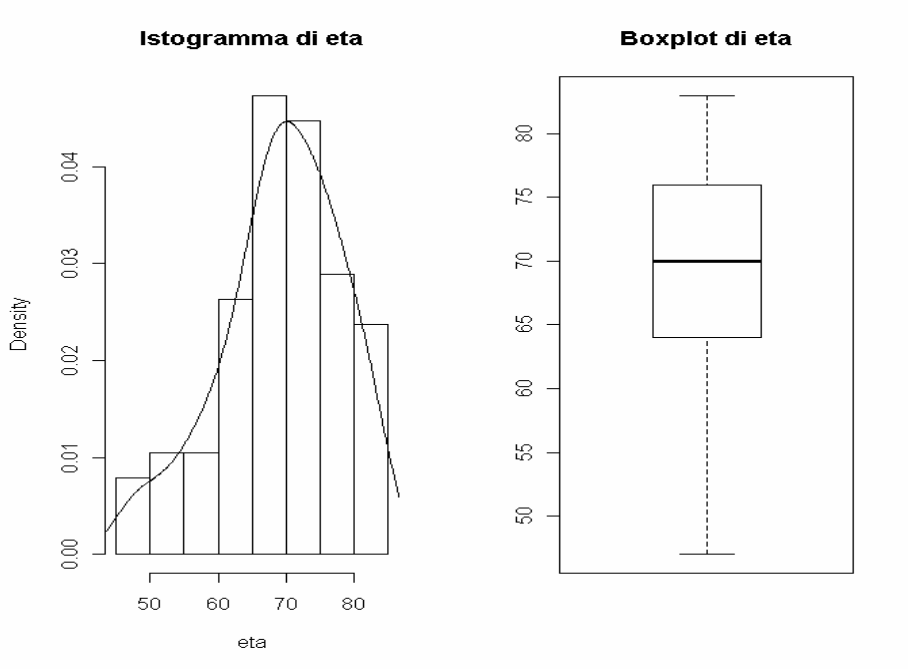

Figura 2.1: Istogramma e boxplot della variabile età.

Dall'istogramma (Figura 2.1) si nota un'asimmetria nella distribuzione dell'età. Il test di Shapiro-Wilk porta al rifiuto dell'ipotesi di normalità (W_{SW} = 0.961; p-value = 0.021).

Si procede quindi esaminando la variabile `eta` per genere, per verificare un'eventuale differenza tra maschi e femmine (Figura 2.2).

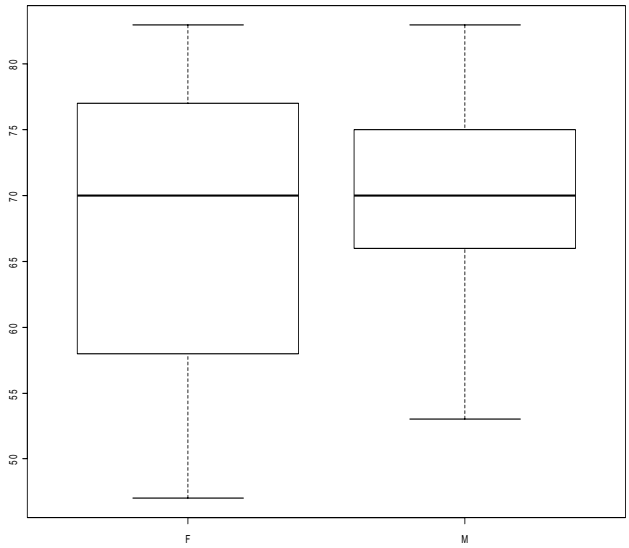

Figura 2.2: Boxplot dell'età distinta per sesso.

L'età media per le femmine è 67.4 anni (sd =11.78), mentre per i maschi è 70.2 anni (sd = 6.89). Il test di Wilcoxon basato sui ranghi porta all'accettazione dell'ipotesi nulla di uguaglianza delle due distribuzioni (`W = 586, p-value = 0.5725`). L'età, al momento della diagnosi, è pertanto omogenea rispetto al genere.

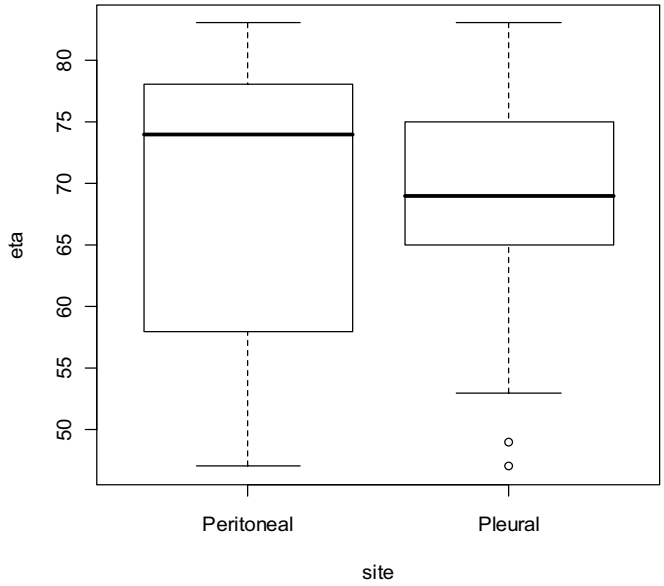

Figura 2.3: Boxplot dell'età distinta per `site`.

I pazienti sono quindi suddivisi in 76% casi di mesotelioma pleurico e 24% peritoneo. Il boxplot della variabile `eta` (Figura 2.3), secondo la suddivisione dei pazienti per localizzazione del tumore riscontrata (`peritoneal` o `pleural`), evidenzia che i due mesoteliomi sembrano differenti per variabilità. Il test di Wilcoxon porta all'accettazione dell'ipotesi nulla di uguaglianza delle due distribuzioni (`W = 574.5, p-value = 0.525`).

I pazienti vengono inoltre distinti secondo le tre sottocategorie di `subtype`, (`epithelioid`, `mixed` e `sarcomatous`), in relazione al tipo di cellula maligna presente nel tumore (Figura 2.4).

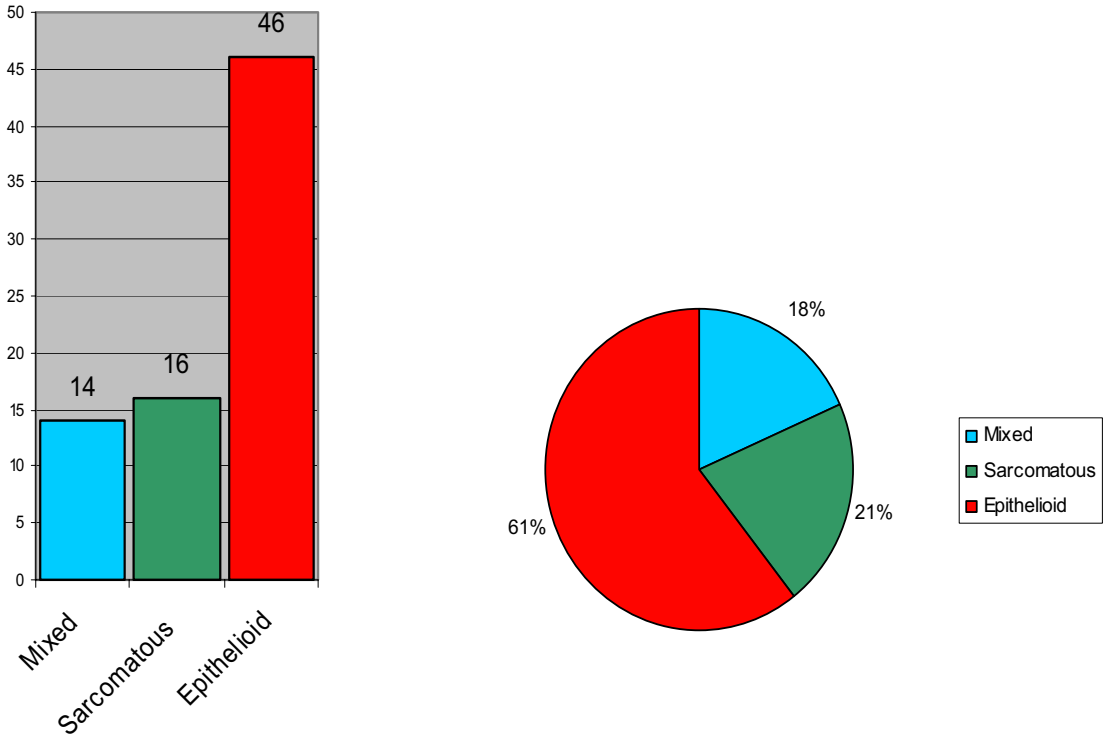

Figura 2.4: Diagramma a barre e diagramma a torta per `subtype`.

Il MM più diffuso, con il 61% di casi, è l'epitelioide, mentre il misto e sarcomatoso sono rilevati nel 18% e 21% di casi, rispettivamente (Figura 2.4).

I pazienti a cui è stato diagnosticato il MM sottotipo `mixed` ha un'età media pari a 71.0 anni (sd = 8.99), per `sarcomatous` l'età media è uguale a 72.2 anni (sd = 7.23), mentre l'età media per `epithelioid` è 67.74 anni (sd = 9.03).

La distribuzione dell'età per `subtype` è rappresentata in Figura 2.5. Attraverso il test di Kruskal-Wallis, si può concludere che la variabile `subtype`, rispetto l'età, non si differenzia in distribuzione nei i tre livelli di sottoclassificazione istologica del MM (`Kruskal-Wallis chi-squared = 28.81, df = 29, p-value = 0.475`).

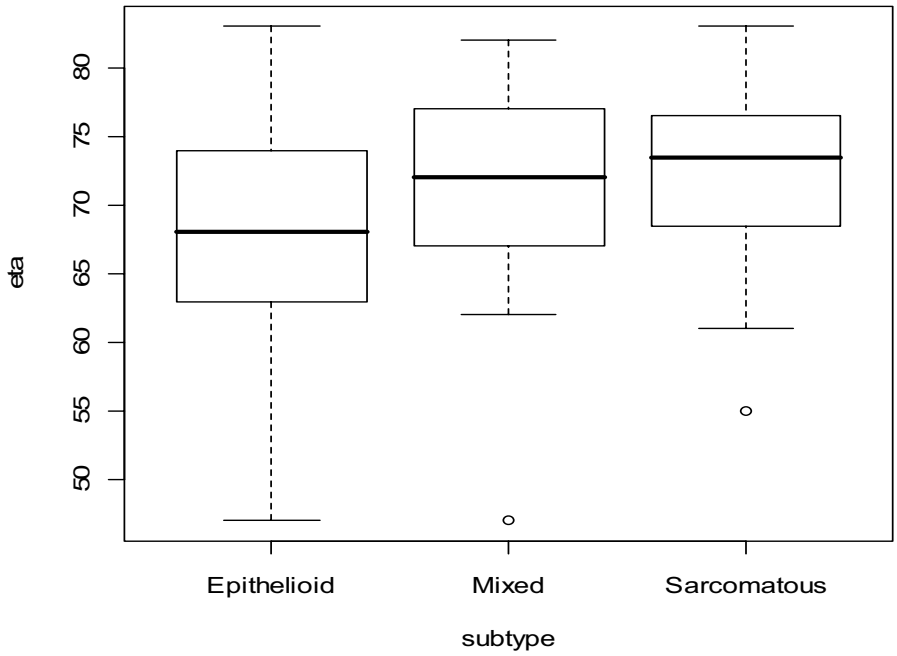

Figura 2.5: Boxplot dell'età in funzione di `subtype`.

La variabile dicotomica `status` evidenzia l'elevato tasso di mortalità riscontrato nei pazienti a cui viene diagnosticato il MM: al 31 dicembre 2010, risultano in vita solo il 22% dei pazienti (ovvero 17 pazienti su 76). A causa di questa censura, nel prossimo capitolo si condurrà l'analisi della sopravvivenza, tenendo conto dei dati censurati.
La statistica test χ^2 per la verifica dell'indipendenza tra la variabile `status` e le altre variabili categoriali accetta l'ipotesi nulla in tutti e quattro i casi (Tabella 2.3).

Variabile	*gl*	χ^2	*p-value*
`subtype`	2	1.014	0.602
`site`	1	0.116	0.733
`sesso`	1	0.409	0.522
`eta`	29	26.859	0.579

Tabella 2.3: Risultati dei test χ^2 per l'indipendenza di `status`.

La variabile di durata `survival` indica i mesi di sopravvivenza calcolati dalla diagnosi del tumore, e ha media di circa 20 mesi (sd = 22.12). La mediana di sopravvivenza è pari a 11 mesi, anche se si deve sempre tener presente che alcuni dati sono censurati e quindi non è noto quando avviene l'evento finale, in questo caso il decesso, rispetto alla diagnosi del MM.

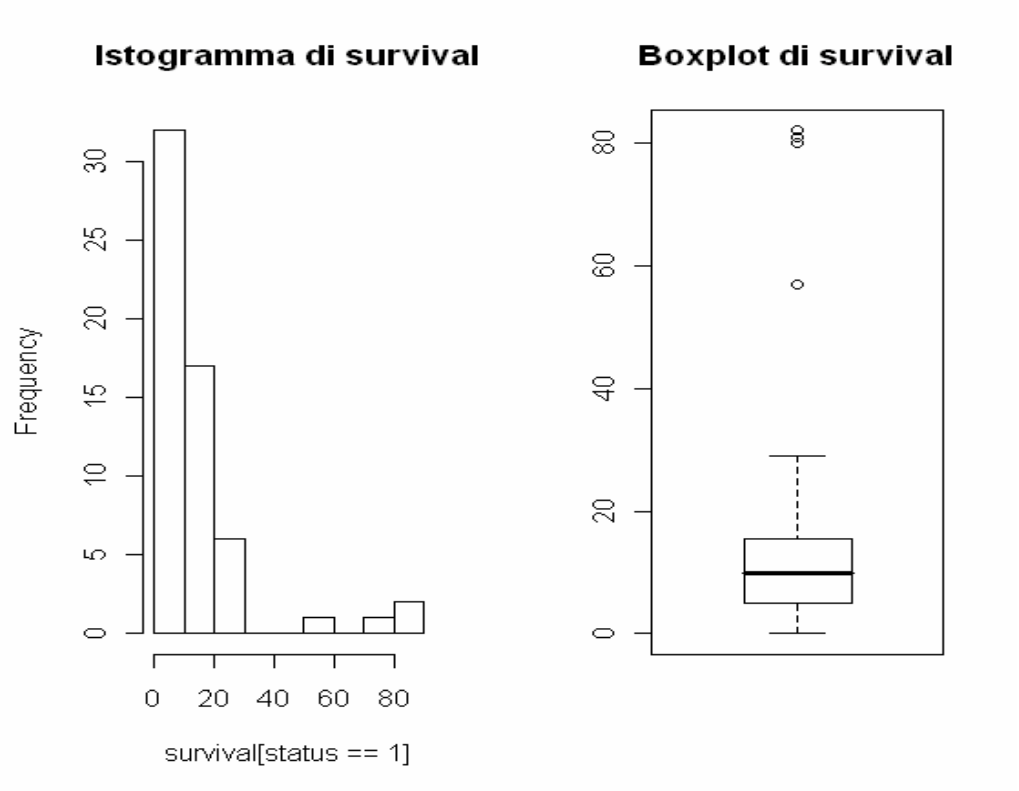

Figura 2.6: Istogramma e boxplot di `survival` per dati non censurati.

Nelle Figure 2.6 e 2.7 sono riportati i grafici relativi ai tempi non censurati (78% dei casi in esame). Dalle Figure 2.6, si nota l'asimmetria positiva, tipica delle variabili di durata. La Figura 2.7 riassume, considerando i cinque periodi temporali ritenuti più rappresentativi, i mesi di sopravvivenza dal riscontro del MM: 21 pazienti (35%) sono sopravvissuti meno di 6 mesi; mentre in 4 casi (7%) oltre i 36 mesi dalla diagnosi.

Si fa presente che per 5 casi, i mesi di sopravvivenza sono 0 in quanto il paziente è deceduto poco dopo la diagnosi del MM, questo a causa dello stato già avanzato del tumore.

Non è presente una correlazione significativa tra le due variabili `eta` e `survival` (`cor = 0.061`).

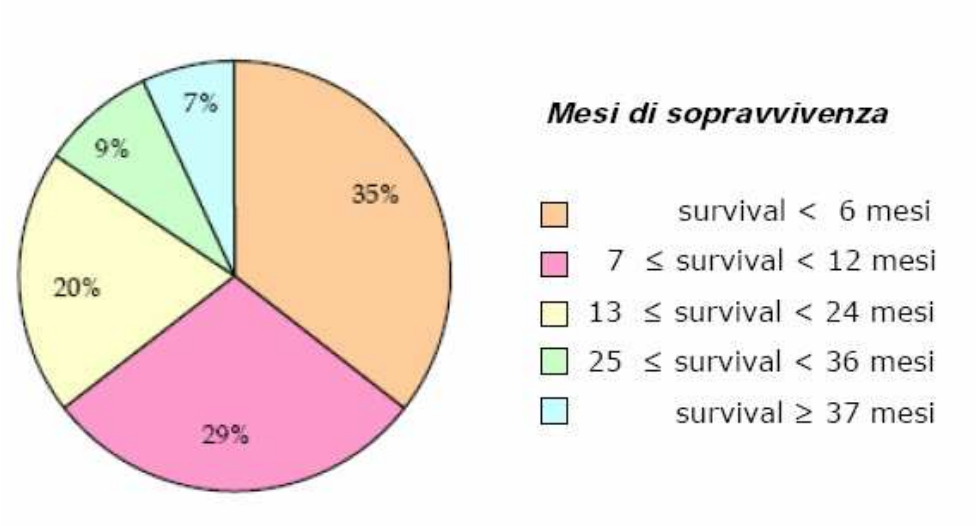

Figura 2.7: Diagramma a torta dei mesi di sopravvivenza per classi.

Per la classificazione della variabile `survival` è d'interesse verificare l'ipotesi nulla di indipendenza con le variabili categoriali `site` e `subtype`. A tale scopo si considera la statistica χ^2. I risultati ottenuti sono riportati in Tabella 2.4: la variabile `survival` è dipendente da `subtype`, mentre è indipendente sia da `site` che da `sesso`.

Variabile	*gl*	χ^2	*p-value*
`subtype`	8	26.220	0.001
`site`	4	4.754	0.313
`sesso`	4	1.689	0.793

Tabella 2.4: Risultati dai test χ^2 per l'indipendenza di `survival`.

Si conclude quindi con l'analisi dei risultati ottenuti dall'immunoistochimica *IHC* che, come citato al §1.3, permette di determinare l'espressione dei marcatori della transizione epiteliale - mesenchimale (EMT), evidenziando l'espressione e la localizzazione della proteine *E-caderina, N-caderina, β-catenina, citocheratine* 5/6,

vimentina, αSMA*, S100A4 e MMP-2 e MMP-9*, in una sezione di tessuto dei pazienti in osservazione.

Confrontando il boxplot per ogni proteina (Figura 2.8), si nota che hanno comportamenti simili:

- *N-caderina* con *β-cadenina* e `S100A`;
- *citocheratine 5/6* con `α-SMA` e `mmp2` e `zeb2`;
- *vimentina* e `zeb1`;
- *E-caderina* con `mmp9`.

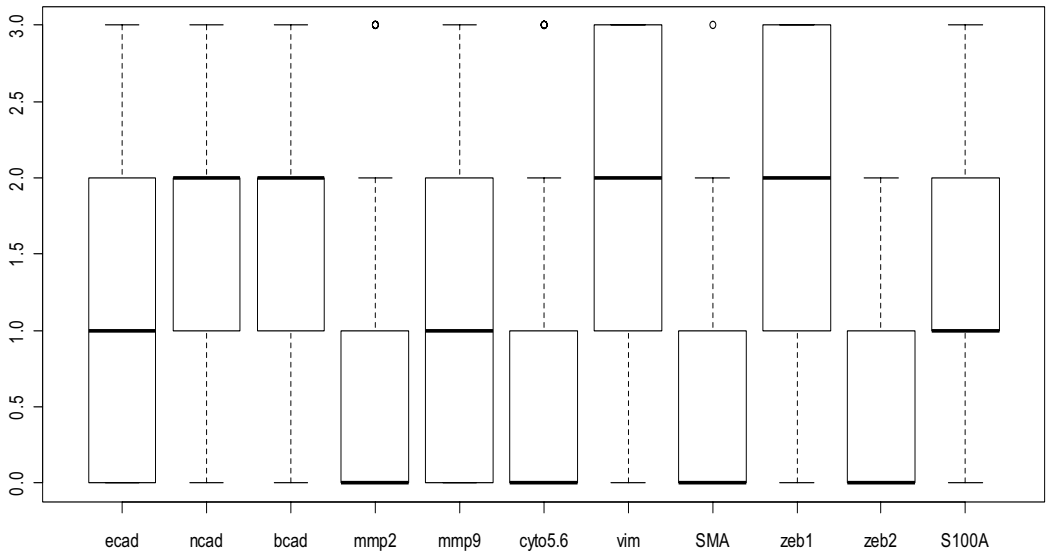

Figura 2.8: Boxplot delle varie proteine esaminate.

Dal calcolo del coefficiente di correlazione di *Spearman* tra le varie proteine (Tabella 2.5) non emergono correlazioni significative. L'unico indice di correlazione positivo significativo ($\rho = 0.64$) risulta essere tra `ecad` e `bcad`. Tra i più rilevanti, si possono citare anche l'indice di correlazione positivo tra `vim` e `SMA` ($\rho = 0.53$), mentre tra `zeb1` ed `ecad` si ottiene invece un indice di correlazione negativo significativo ($\rho = -0.54$).

	ecad	ncad	bcad	mmp2	mmp9	cyto5.6	vim	SMA	zeb1	zeb2	S100A
ecad	1.000										
ncad	0.163	1.000									
bcad	0.636	0.494	1.000								
mmp2	-0.190	0.152	0.022	1.000							
mmp9	-0.133	0.077	-0.056	0.290	1.000						
cyto5.6	0.263	-0.233	0.122	-0.148	-0.121	1.000					
vim	-0.368	0.238	-0.210	0.233	0.174	-0.359	1.000				
SMA	-0.451	0.073	-0.366	0.192	0.139	-0.167	0.545	1.000			
zeb1	-0.537	0.129	-0.334	0.246	0.175	-0.078	0.529	0.415	1.000		
zeb2	-0.392	-0.056	-0.306	0.166	0.069	-0.079	0.312	0.476	0.286	1.000	
S100A	-0.214	0.175	-0.204	0.049	0.211	-0.266	0.530	0.475	0.144	0.338	1.000

Tabella 2.5: Coefficiente di correlazione di *Spearman* calcolato tra le variabili proteine.

In Figura 2.9, per ogni proteina si riporta il boxplot ottenuto sia in funzione della variabile `subtype` (a sinistra di ogni riquadro) che di `site`.

In particolare, osservando i boxplots per `subtype` (a sinistra di ogni riquadro) si nota che ogni proteina è complessivamente presente nelle tre sottocategorie, a distinzione di qualche caso come per esempio, le proteine *E-caderina*, *citocheratine 5/6* e *vimentina* risultano praticamente assenti per il sottotipo `Sarcomatous` mentre, per il sottotipo `Ephitelioid`, non vi è particolare presenza delle proteine α-SMA e `S100A`.

Per quanto riguarda i boxplots relativi a `site` (a destra di ogni riquadro) si nota che, per ogni proteina i boxplot per i le due tipologie `pleural` e `peritoneal` sono simili.

Per la tipologia `pleural`, sembrerebbe che le tre *caderine* (`ecad`, `ncad`, `bcad`) abbiano un ruolo maggiore rispetto `peritoneal` che, invece risulta particolarmente influenzata da `mmp9`.

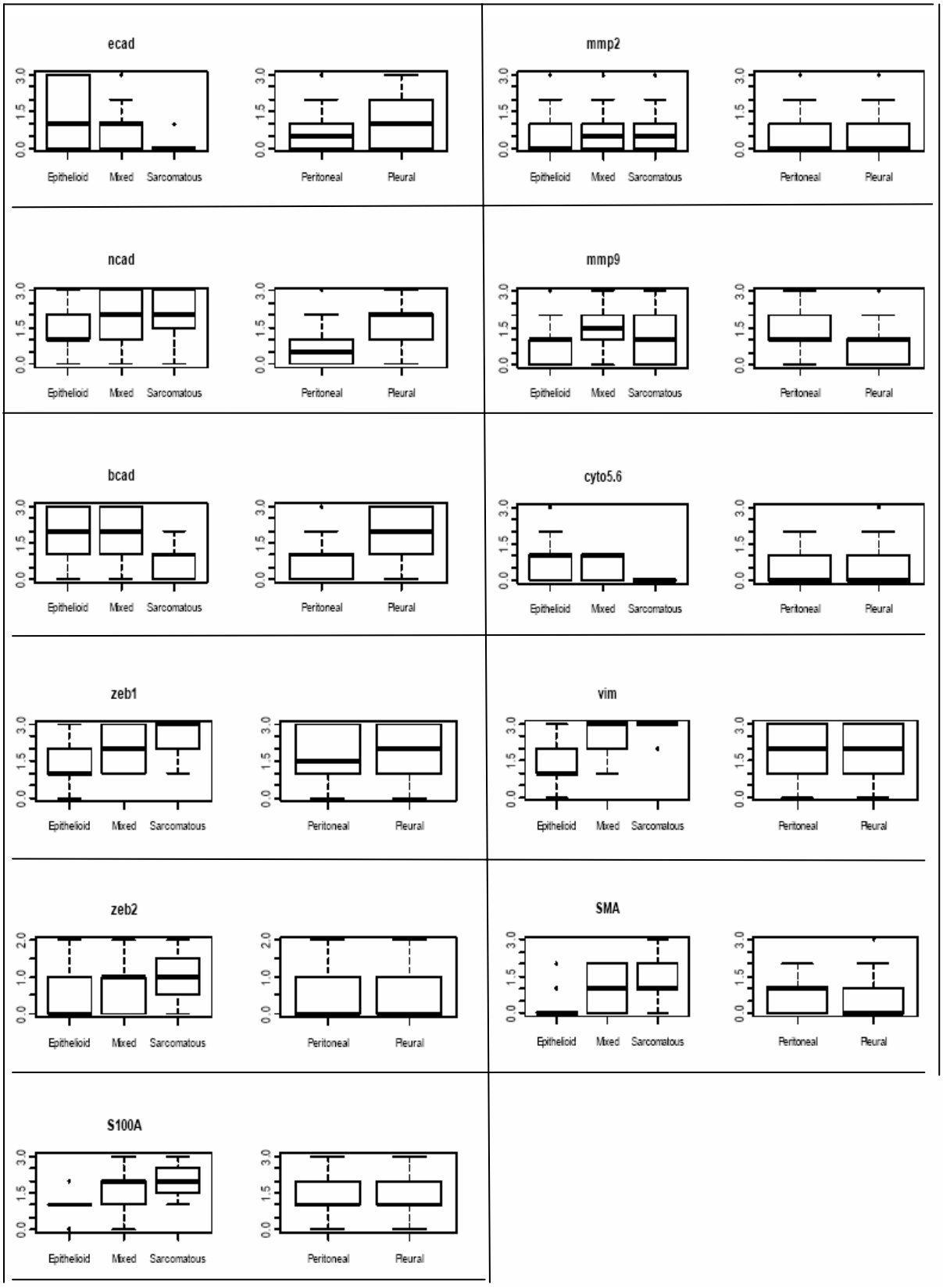

Figura 2.9: Boxplot delle proteine in funzione di `subtype` e `site`.

2.2 Considerazioni finali

Dall'analisi esplorativa delle variabili del *dataset,* si sono ottenute diverse informazioni in merito alle varie caratteristiche rilevate sui 76 pazienti a cui è stato diagnosticato il MM.

Nel prossimo capitolo si effettuerà lo studio sulla variabile `survival` per esaminare quali fattori possono influenzarla; in particolare, ci si aspetta che sia l'età che il genere del paziente risultino ininfluenti per l'analisi della sopravvivenza.

CAPITOLO 3

ANALISI DELLA SOPRAVVIVENZA

3.1 Introduzione

In vari settori applicativi si pone il problema di analizzare dati che rappresentano, per ciascuna unità, il tempo trascorso, dall'inizio dell'osservazione, fino al verificarsi di un evento d'interesse. In Ingegneria questo studio è noto come "teoria dell'affidabilità", mentre in Economia o in Sociologia è chiamato "analisi di durata"; nelle scienze biomediche, viene definito come "analisi della sopravvivenza". Per un'introduzione si veda Pace e Salvan (2001, Cap.11).

Questo tipo di analisi si adatta bene alle situazioni in cui il problema generale è valutare la probabilità di sopravvivenza in funzione del tempo, eventualmente in dipendenza da altre variabili in studio. L'analisi della sopravvivenza è costituita, infatti, dalla presenza di una variabile aleatoria non negativa, con distribuzione tipicamente asimmetrica, legata al tempo di accadimento di un particolare evento, nel caso in esame il decesso del paziente. Per maggiori approfondimenti si veda Marubini e Valsecchi (1995).

Per definizione, la funzione di sopravvivenza $S_T(t)$ esprime la probabilità che il tempo di sopravvivenza T dell'unità sperimentale sia maggiore di t, ossia

$$S_T(t) = 1 - F_T(t) = P(T > t),$$

con $F_T(t)$ funzione di ripartizione di T.

I modelli per l'analisi della sopravvivenza consentono di valutare la relazione fra fattori prognostici e il tempo di sopravvivenza e sono generalmente espressi in termini della funzione di azzardo $\lambda(t)$ (o *hazard function*), definita come

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{\Pr\{t \le T < t + \Delta t \mid T \ge t\}}{\Delta t} = \frac{f_T(t)}{S_T(t)},$$

con $f_T(t)$ densità di probabilità della variabile casuale T che descrive i tempi di sopravvivenza dei pazienti.

Nell'ambito dei dati di durata, la funzione di azzardo $\lambda(t)$ assume un ruolo cruciale in quanto fornisce indicazioni sul rischio istantaneo che l'evento si verifichi nell'intervallo $[t, t + \Delta t]$, condizionatamente al fatto che il soggetto è vivo al tempo a t.

Quindi si può definire la funzione cumulativa di rischio, convenzionalmente indicata Λ, come

$$\Lambda = - \log S(t)$$

così, $\frac{d}{dt}\Lambda(t) = -\frac{S'(t)}{S(t)} = \lambda(t)$.

Le precedenti definizioni implicano $\Lambda(t) = \int_0^t \lambda(u)\,du$, ovvero l'accumulazione di rischio nel tempo, con $\lambda(u)$ generica stima del rischio integrato.

L'obiettivo dell'analisi della sopravvivenza può essere:

- stimare e interpretare funzioni di sopravvivenza e l'azzardo a partire da dati di sopravvivenza;
- confrontare funzioni di sopravvivenza e l'azzardo tra gruppi di soggetti con caratteristiche diverse;
- valutare l'impatto di variabili tempo-indipendenti e tempo-dipendenti sulla sopravvivenza.

Molto diffuso, nell'analisi di sopravvivenza, è il problema di osservazioni mancanti o incomplete, ovvero di dati censurati. Gli schemi di censura più ricorrenti (Klein e Moeschberger, 2003), possono essere raggruppati sostanzialmente in tre classi:

1. *Censura 1° tipo*: i soggetti sono osservati per un periodo di tempo fissato. Alla fine dello studio i soggetti che non presentano fallimento risultano censurati;
2. *Censura 2° tipo*: si differenzia dal 1° tipo, in quanto il numero totale di fallimenti è stabilito a priori. La lunghezza dello studio quindi non risulta fissata;
3. *Censura casuale*: il totale del periodo di osservazione è fissato, ma i soggetti entrano in studio in tempi differenti. Alcuni individui falliscono, altri individui risultano persi dal *follow-up*, altri ancora non presentano fallimento alla fine dello studio.

La bontà delle stime è quindi relazionata al numero di eventi piuttosto che al numero di osservazioni: maggiore è il numero di valori non censurati migliori saranno le stime dei coefficienti.

Nel caso del MM, la censura è da considerarsi casuale. Alcuni soggetti (in totale 17) "sopravvivono" oltre il tempo di osservazione e quindi, non conoscendo in quale particolare istante futuro essi andranno incontro al "decesso", si parla di dati troncati o censurati, e sono associabili allo "`status=0`".

Per i dati non censurati, ovvero i 59 pazienti deceduti al 31/12/2010, si assume lo "`status=1`", perché si considera la loro storia clinica completa.

3.1.1 Modelli parametrici vs non parametrici

La scelta della distribuzione di sopravvivenza esprime alcune particolari informazioni sulla relazione del tempo su qualsiasi variabile esogena riguardo alla sopravvivenza.
È naturale scegliere una distribuzione statistica che non ha supporto negativo, in quanto i tempi di sopravvivenza sono per l'appunto positivi.
Inoltre possono essere usati metodi parametrici o non parametrici.
I modelli parametrici richiedono che la distribuzione del tempo di sopravvivenza sia nota e la funzione d'azzardo sia completamente specificata, ad esclusione dei valori di alcuni parametri. Ci sono diverse distribuzioni e le più comuni, usate nell'analisi dei tempi di sopravvivenza, prevedono funzioni di densità quali l'e*sponenziale* con funzione di rischio costante, la *Gamma*, la *Weibull* con funzione di rischio monotona e la l*og-normale* con funzione di rischio crescente sino ad un massimo e poi decrescente (Figura 3.1).

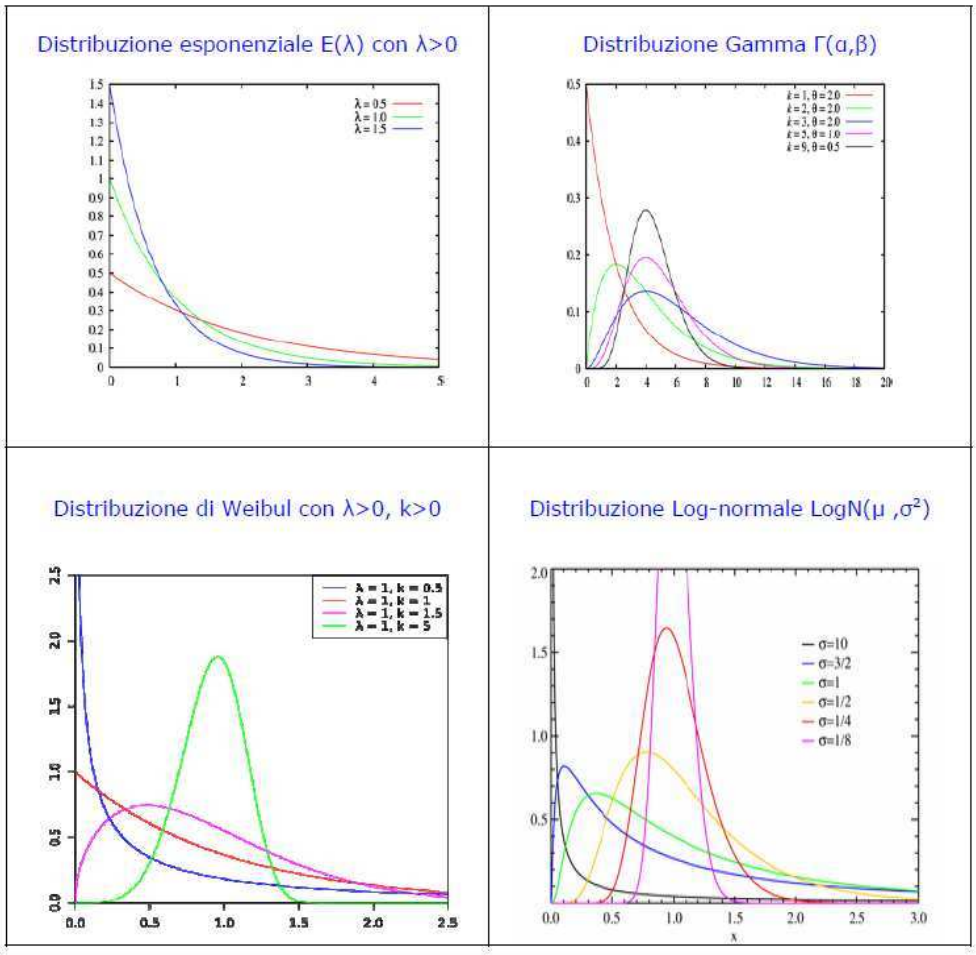

Figura 3.1: Distribuzioni più utilizzate in ambito parametrico.

Quando le assunzioni che sono alla base dei modelli parametrici classici non sono verificate, si utilizzano i metodi non parametrici, che per l'appunto non necessitano di assunzioni sulla distribuzione di *T*.

Lo strumento di analisi statistica non paramatrica che consente di costruire le curve di sopravvivenza (ovvero il grafico della probabilità di sopravvivenza) e di misurare il rischio osservato è il metodo di *Kaplan-Meier* in cui la funzione di azzardo non è specificata.

Quando vi è difficoltà a formulare specifiche ipotesi per individuare un modello che rappresenti la distribuzione dei tempi di sopravvivenza, è infine possibile ricorre anche al modello semi-parametrico di *Cox* a rischi proporzionali, con eventuale opzione di stratificazione dei dati in sottogruppi (es. sesso, patologia etc.). Quest'ultimo è meno efficiente dei modelli parametrici, ma il suo utilizzo è giustificato quando (Siegel e Castellan, 1988) i dati non si conformano al tipo di distribuzione richiesto dalle procedure parametriche.

3.2 Lo stimatore di Kaplan-Meier

Lo stimatore non parametrico della curva di sopravvivenza più comune è quello di *Kaplan-Meier* (K-M).

Considerati i tempi distinti e ordinati, relativi ad eventi accaduti tra gli *n* soggetti in esame, avremo un ordinamento del tipo: $t_1 < t_2 < ... < t_J$, con $J \leq n$.

Sia quindi d_j il numero di decessi che avvengono al tempo t_j *e sia* n_j il numero di soggetti a rischio al tempo t_j ($j=1,\ldots,J$).

La probabilità p_j di sopravvivere oltre il tempo t_j, condizionatamente all'essere sopravvissuti fino all'istante precedente a t_j , è stimata da

$$\hat{p}_j = \frac{n_j - d_j}{n_j} = 1 - \hat{q}_j ,$$

dove $\hat{q}_j = \frac{d_j}{n_j}$ è la stima della probabilità condizionata di subire l'evento al tempo t_j, con $j= 1,\ldots,J$.

La funzione di sopravvivenza che viene quindi stimata è un prodotto di probabilità di sopravvivenza, data da

$$\widehat{S}_{KM}(t) = \prod_{t_j \leq t} \frac{n_j - d_j}{n_j} ,$$

che cambia valore solamente quando si verifica almeno un evento.

Lo stimatore di K-M risulta non distorto a varianza minima ed è caratterizzato da una distribuzione asintotica gaussiana con varianza stimabile tramite la formula di Greenwood (1926), data da

$$\hat{V}ar\left[\hat{S}_{KM}(t)\right] = \hat{S}_{KM}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}.$$

Intervalli di confidenza per la funzione di sopravvivenza di livello approssimato (1-α), sono calcolabili (Sawyer, 2003) come

$$\widehat{S}_{KM}(t) \pm z_{\alpha/2}\sqrt{\hat{V}ar\left[\hat{S}_{KM}(t)\right]},$$

con $z_{\alpha/2}$ quantile di livello α/2 della distribuzione normale standard.

Per considerare valido il modello di K-M, devono essere rispettate alcune assunzioni:

- censura indipendente dal gruppo di classificazione;
- presenza di un numero limitato di dati *censored*: si deve infatti tenere presente che questi ultimi influiscono sulla stima della curva di K-M dato che diminuisce il numero di pazienti a rischio, rendendo la stima di sopravvivenza meno precisa di quanto non si avrebbe in presenza di un numero ridotto di dati censurati;
- campione di dimensione sufficientemente grande, così le curve di K-M hanno maggior precisione.

Graficamente (si vedano le Figure 3.2 e 3.3), la stima di Kaplan-Meier è una curva a gradini continua che parte da 1 (infatti $\widehat{S}_{KM}(t_0) = 1$) e decresce nel tempo; ha una caduta in ogni istante t_j in cui si verifica almeno un decesso; gli istanti temporali in corrispondenza dei quali le osservazioni sono censurate, ovvero un paziente esce dallo studio, vengono riportati nel grafico con un segno verticale o il simbolo " + ".

Come la funzione di ripartizione empirica, anche la stima di K-M è costante a tratti: parte da un livello di sopravvivenza pari al 100% (infatti, per definizione, tutti i pazienti sono vivi al tempo "0", che coincide con il momento della diagnosi) e decresce nel tempo fino a tendere allo zero.

La Figura 3.2 rappresenta la stima della curva di sopravvivenza di K-M in relazione alla classificazione tumorale "`site`": la curva rossa si riferisce a pazienti con `site=Peritoneal`, mentre la curva tratteggiata corrisponde ai pazienti con `site=Pleural`. Al contrario della funzione di colore rosso, la curva nera presenta molti tratti di linea costante, indice che in quel periodo di tempo non si sono verificati decessi.

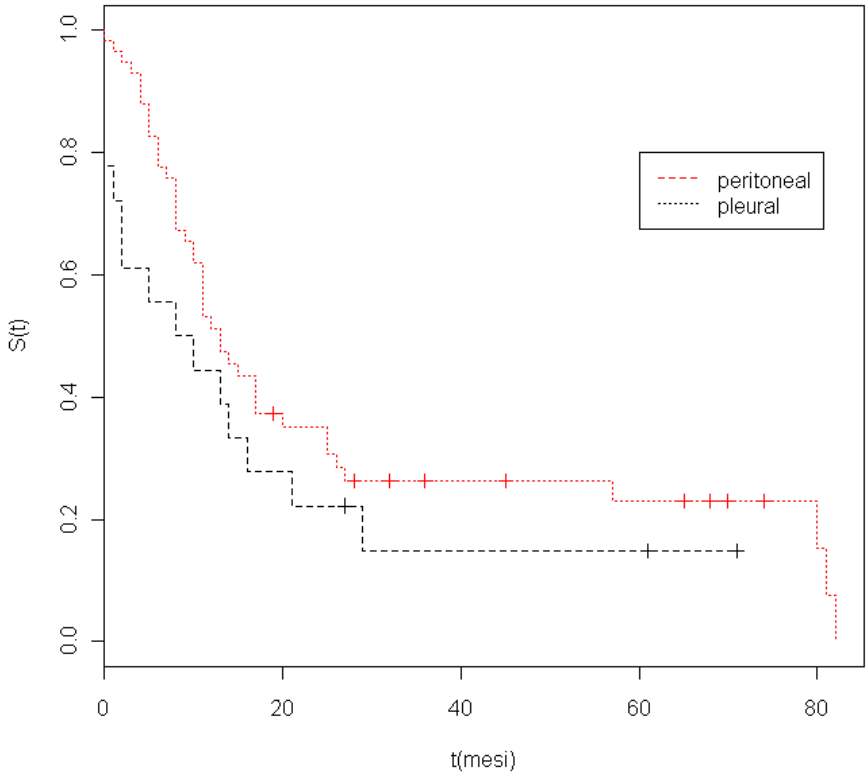

Figura 3.2: Curve di sopravvivenza di Kaplan-Meier per i 2 gruppi `site`.

Spesso viene usato come indicatore il tempo mediano di sopravvivenza. In presenza di dati censurati, è facile da esprimere se tutti i tempi censurati sono superiori alla mediana (non si potrebbe calcolare la media aritmetica): se la mediana riguarda un intervallo di tempo, si prende il tempo centrale altrimenti, si considera il tempo t per cui $Pr(T > t) = 0.50$.

Per `site=Peritoneal` il tempo di sopravvivenza mediano risulta pari a 9 mesi (95% I.C. = (2 , 29) mesi), mentre per il `site=Pleural` il tempo di sopravvivenza mediano è 13 (95% I.C. = (11 , 25) mesi).

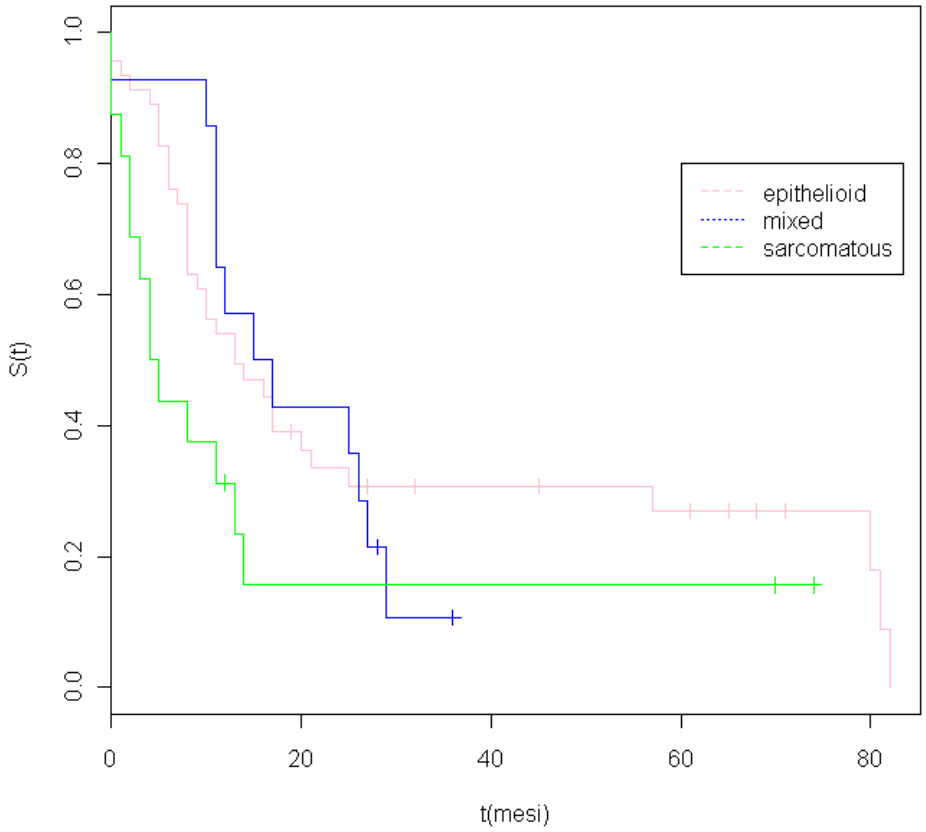

Figura 3.3: Curve di Kaplan - Meier per i 3 gruppi `subtype`.

La Figura 3.3 rappresenta invece le curve di sopravvivenza in relazione a `subtype`: la curva di colore rosa si riferisce al sottotipo epitelioide che, rispetto alle altre due, ha sopravvivenza più lunga; poi vi è la categoria `Sarcomatous` (colore verde) e la mista con durata minore di tutte (inferiore ai 40 mesi).

I tre tempi di sopravvivenza mediani calcolati per per `subtype` sono: 13.0 mesi per `Epithelioid`, 16.0 mesi per `Mixed` e 4.5 mesi per `Sarcomatous`.

Per valutare la differenza fra due o più curve di sopravvivenza si può ricorrere al *log-rank test* (detto anche test di Mantel-Haenszel). Tale test, basato sui ranghi, è da utilizzare quando non è ipotizzata una distribuzione dei tempi all'evento per i gruppi di soggetti in esame, per saggiare l'ipotesi nulla

$$H_0 : S_A(t) = S_B(t), \ \forall t.$$

La statistica *log-rank* test è costruita a partire da una serie ordinata di tabelle di contingenza 2x2, una per ognuno dei *J* tempi dell'evento.
La tabella generica, relativa all'*j*-mo tempo è rappresentata come nella Tabella 3.1.

Gruppo	*Morti al tempo* t_j	*Vivi al tempo* t_j	*Soggetti a rischio appena prima di* t_j
A	d_{Aj}	n_{Aj}-d_{Aj}	n_{Aj}
B	d_{Bj}	n_{Bj}-d_{Bj}	n_{Bj}
totale	d_j	n_j-d_j	n_j

Tabella 3.1: Tabella di contingenza 2x2 al tempo dell'evento t_j.

Poiché, per ogni tabella, gli eventi si distribuiscono nei due gruppi proporzionalmente al numero di soggetti ancora a rischio, è possibile calcolare il numero di eventi attesi. Un eventuale scostamento, tra il totale degli eventi osservati e quelli attesi in ciascun gruppo, suggerisce un diverso livello di mortalità nei due gruppi (Grigoletto, 2010).
Sotto l'ipotesi nulla la distribuzione dei casi risulta ipergeometrica, per cui il valore atteso (condizionato) di d_{Aj} è

$$E(d_{Aj}) = \left[n_{Aj} \times \frac{d_j}{n_j} \right],$$

mentre la varianza (condizionata) è

$$\text{var}(d_{Aj}) = \left[n_{Aj} \times \frac{d_j}{n_j} \left(1 - \frac{d_j}{n_j} \right) \right] \left(\frac{n_j - n_{Aj}}{n_j - 1} \right),$$

la cui formula deriva dal prodotto di due termini: il primo, tra parentesi quadre, è la stima della varianza di una variabile casuale binomiale, il secondo è il fattore di correzione per il campionamento da una popolazione finita di dimensione n_j.
Sotto ipotesi nulla si dimostra che il rapporto

$$Q_{M-H} = \frac{\left\{ \sum_{j=1}^{J} \left[d_{Aj} - E(d_{Aj}) \right] \right\}^2}{\sum_{j=1}^{J} \text{var}(d_{Aj})}$$

si distribuisce asintoticamente come un χ^2 con 1 grado di libertà.
Quanto sopra descritto è talvolta esteso anche per il confronto tra tre o più gruppi.
Nella Tabella 3.2, vengono riepilogati i risultati ottenuti effettuando il *log-rank test* per le due variabili `site` e `subtype`: in entrambi i casi si accetta l'ipotesi nulla. Il *p-values* ottenuti indicano infatti che non vi sia distinzione tra i due `site` (p = 0.186) e le tre classificazioni del mesotelioma (p = 0.099).

Variabile		*N° oss. per gruppo*	*varianza*	*Log-rank test* (Q_{M-H})	*p-value*
site	Peritoneal	18	1.75	1.70	0.186
	Pleural	58	1.75		
subtype	Epithelioid	46	1.56	4.60	0.099
	Mixed	14	0.10		
	Sarcomatous	16	4.61		

Tabella 3.2: *Log-rank test* calcolato per le variabili `site` e `subtype`.

Il *log-rank* test è utile se si vuole paragonare l'effetto che un singolo fattore di rischio ha sulla sopravvivenza, ma talvolta risulta limitato quando i fattori in studio sono più di uno e si vogliono valutare contemporaneamente. In tali casi è più utile ricorrere al modello di regressione di Cox (Cox, 1972).

3.3 Il modello di Cox

L'interesse del ricercatore che si occupa dell'analisi dei tempi di sopravvivenza non è solo rivolto alla stima della funzione di sopravvivenza o della funzione d'azzardo, ma anche al confronto dell'esperienza di vita di due o più insiemi di individui che differiscono tra loro per una certa caratteristica. Inoltre, in campo medico, si è in generale interessati all'individuazione di fattori prognostici che spieghino nel modo più adeguato eventuali differenze significative nell'esperienza di vita di diversi gruppi di pazienti.
Nei modelli di regressione, la funzione di rischio, che rappresenta la probabilità di morire al tempo t per un individuo con vettore di covariate x sopravvissuto fino al tempo t, viene generalmente fattorizzata in due parti, come

$$\lambda(t;x) = \lambda_0(t) \times H(x,\beta),$$

dove $\lambda_0(t)$ è l'azzardo di base che viene assunto come livello di rischio di riferimento al tempo t (Shoenfeld, 1982), non ha una forma specifica ed è lo stesso per tutti i soggetti. Invece, $H(x,\beta)$ è una funzione delle covariate, ma non del tempo, mentre i β sono i coefficienti associati alle covariate. L'intercetta non è presente poiché è "assorbita" dal rischio di base. L'effetto delle covariate può solo indurre traslazioni proporzionali sul rischio ma non può modificarne l'andamento.

La forma scelta da Cox (1972) per esplicitare l'effetto delle covariate sul rischio di base è quella esponenziale, in cui $H(x,\beta)=exp(\beta^T x)$. Quindi, supposto che vi siano p variabili concomitanti $(x_1,\ldots, x_p)$, il modello proposto da Cox esprime la funzione di azzardo in funzione del tempo e delle covariate, come

$$\lambda(t;x_i) = \lambda_0(t)\exp(\beta^T x_i) = \lambda_0(t)\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}),$$

dove x_{ir} è il valore assunto dalla r-esima variabile concomitante per l'i-esima unità $(i=1,\ldots,n)$ e $(\beta_1,\ldots,\beta_p)$ sono i p parametri ignoti di regressione.

Caratteristica particolare di questo modello è permettere di fare inferenza sul vettore dei parametri di regressione β senza richiedere necessariamente la specificazione di una classe parametrica per $\lambda_0(t)$. Inoltre, i predittori non dipendono dall'istante t in cui viene fatta la valutazione: il rapporto di rischio è quindi lo stesso indipendentemente dal tempo (Bland, 2009).

La stima dei parametri incogniti β del modello e le relative procedure di verifica di ipotesi avvengono tramite il metodo della massima verosimiglianza parziale (Cox, 1972) che consente di introdurre esplicitamente il troncamento.

Di solito la stima dei parametri β si ottiene massimizzando la verosimiglianza parziale

$$L_P(\beta) = \prod_{i=1}^{n}\left[\frac{\exp(x_i^T\beta)}{\sum_{j\geq i}\exp(x_j^T\beta)}\right]^{\delta_i},$$

dove δ_i indica il numero degli individui per i quali si registra l'evento nel tempo t_i, o nel modo equivalente, risolvendo l'equazione di verosimiglianza parziale

$$\sum_{i=1}^{n}\delta_i\left[x_i - \frac{\sum_{j\geq i}\exp(x_j^T\beta)x_j}{\sum_{j\geq i}\exp(x_j^T\beta)}\right] = \sum_{i=1}^{n}\delta_i\left[x_i - \frac{S^{(1)}(t_i;\beta)}{S^{(0)}(t_i;\beta)}\right] = 0,$$

la cui soluzione (si veda anche Heritier *et al.*, 2009), denotata con $\hat{\beta}_{PLE}$, è la stima di massima verosimiglianza parziale, nel seguito abbreviata con *PLE* (*partial likelihood estimator*). In dettaglio, la quantità $\left[x_i - \frac{\sum_{j \geq i} \exp(x_j^T \beta) x_j}{\sum_{j \geq i} \exp(x_j^T \beta)} \right]$ è il risultato ottenuto dalla differenza della covariata x_i con l'azzardo al tempo t_i per il soggetto *i*-esimo sull'azzardo di tutti i soggetti in cui non si verifica l'evento al tempo t_i.

Sotto assunzioni di regolarità, il *PLE* è asintoticamente normale con media β e varianza $V = I(\beta)^{-1}$, tipicamente stimata come

$$\hat{I}(\beta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial U_i}{\partial \beta},$$

dove $U_i = \delta_i \left[x_i - \frac{S^{(1)}(t_i; \beta)}{S^{(0)}(t_i; \beta)} \right]$ è l'*i*-esima componente della funzione *score* parziale.

La distribuzione asintotica può essere utilizzata per testare l'ipotesi nulla $H_0 : \beta_j = 0$. A tale scopo, si definisce la statistica *z*, data da

$$z = \frac{\hat{\beta}_{PLE_j}}{SE(\hat{\beta}_{PLE_j})}, \tag{3.1}$$

dove lo *standard error* di $\hat{\beta}_{PLEj}$ è $SE(\hat{\beta}_{PLEj}) = \sqrt{n^{-1} \left[I(\hat{\beta}_{PLE})^{-1} \right]_{jj}}$.

La statistica *z* si distribuisce come una normale standardizzata.

In riferimento alla teoria della verosimiglianza, vi sono tre statistiche test, asintoticamente equivalenti, per saggiare ipotesi del tipo $H_0 : \beta_{(2)} = \beta_{(2)}^0$, con $\beta_{(1)}$ non specificata e $\beta = (\beta_{(1)}, \beta_{(2)})^T$. In particolare il test alla Wald per verificare l'ipotesi $H_0 : \beta = \beta_{(2)}^0$ è

$$\chi_W^2 = \left[(\hat{\beta}_{PLE\,(2)} - \beta_{(2)}^0)^T \hat{V}^{-1}(\beta)_{(22)} (\hat{\beta}_{PLE\,(2)} - \beta_{(2)}^0) \right], \tag{3.2}$$

che si distribuisce come un χ^2 con p gradi di libertà, dove $p = \dim(\beta_{(2)})$ d); con $\hat{V}(\beta)_{(22)}$ blocco di posizione (2,2) della stima della varianza asintotica $\hat{V} = I(\hat{\beta}_{PLE})^{-1}$. In alternativa, per verificare $H_0 : \beta = \beta_{(2)}^0$, si può usare il test *score* dato da

$$\chi^2_{SC} = U(\beta^0_{(2)})^T I(\beta^0_{(2)})^{-1} U(\beta^0_{(2)}), \tag{3.3}$$

che si distribuisce come un χ^2 con p gradi di libertà (Heritier *et al.*, 2009).

3.4 Analisi

Nell'analisi della variabile `survival` si considerano inizialmente tutte le variabili esplicative e si procede con la stima del modello di Cox in *R*. Nello schema seguente viene riportato il risultato ottenuto:

```
> summary(mod.cox)
Call:
coxph(formula = Surv(survival, status) ~ site + subtype + sesso +
    eta + ecad + ncad + bcad + mmp2 + mmp9 + cyto5.6 + vim +
    SMA + zeb1 + zeb2 + S100A)

  n= 76
                       coef exp(coef)  se(coef)      z  Pr(>|z|)
sitePleural        -0.820237  0.440327  0.412645 -1.988  0.04684 *
subtypeMixed       -0.406513  0.665969  0.520625 -0.781  0.43491
subtypeSarcomatous  0.060025  1.061863  0.587084  0.102  0.91856
sessoM             -0.013202  0.986885  0.317660 -0.042  0.96685
eta                -0.023219  0.977048  0.019836 -1.171  0.24178
ecad                0.214963  1.239816  0.196683  1.093  0.27442
ncad                0.009915  1.009964  0.224027  0.044  0.96470
bcad                0.071595  1.074220  0.246028  0.291  0.77105
mmp2                0.074072  1.076885  0.215465  0.344  0.73101
mmp9               -0.127140  0.880611  0.187806 -0.677  0.49842
cyto5.6            -0.607456  0.544735  0.246428 -2.465  0.01370 *
vim                -0.255479  0.774546  0.223756 -1.142  0.25355
SMA                 0.360745  1.434397  0.260549  1.385  0.16619
zeb1                0.553829  1.739903  0.214369  2.584  0.00978 **
zeb2                0.009836  1.009884  0.256579  0.038  0.96942
S100A               0.142586  1.153253  0.262223  0.544  0.58661
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                   exp(coef) exp(-coef) lower .95 upper .95
sitePleural           0.4403     2.2710    0.1961    0.9886
subtypeMixed          0.6660     1.5016    0.2400    1.8476
subtypeSarcomatous    1.0619     0.9417    0.3360    3.3558
sessoM                0.9869     1.0133    0.5295    1.8393
eta                   0.9770     1.0235    0.9398    1.0158
ecad                  1.2398     0.8066    0.8432    1.8229
ncad                  1.0100     0.9901    0.6510    1.5667
bcad                  1.0742     0.9309    0.6632    1.7399
mmp2                  1.0769     0.9286    0.7059    1.6428
mmp9                  0.8806     1.1356    0.6094    1.2725
cyto5.6               0.5447     1.8358    0.3361    0.8830
vim                   0.7745     1.2911    0.4996    1.2009
SMA                   1.4344     0.6972    0.8608    2.3903
zeb1                  1.7399     0.5747    1.1430    2.6485
zeb2                  1.0099     0.9902    0.6108    1.6698
S100A                 1.1533     0.8671    0.6898    1.9281

Rsquare= 0.239   (max possible= 0.996 )

Likelihood ratio test= 20.79  on 16 df,   p=0.1866
Wald test            = 18.67  on 16 df,   p=0.2863
Score (logrank) test = 19.94  on 16 df,   p=0.2229
```

Nell'output, si trovano le stime *PLE* dei coefficienti β, il loro valore esponenziale, l'errore standard e la significatività. Si nota che solamente i coefficienti di `zeb1` e `cyto5.6` e `sitePleural` sono significativi contro l'ipotesi nulla. Inoltre, sono riportati gli intervalli di confidenza al 95% per ogni singolo parametro `[lower .95; upper .95]`e le statistiche (3.2) e (3.3), presentate nel paragrafo precedente.
Come intuito nell'analisi preliminare delle variabili al Capitolo 2, la relazione tra la sopravvivenza e il genere e l'età non risulta significativa.

Poichè nel modello `mod.cox` si ottengono coefficienti non significativi `(Pr(>|z|)> 0.05`, si procede con la regressione all'indietro: ovvero, dal modello appena stimato contenente tutti i possibili predittori, si eliminano, una alla volta, le variabili non significativa, si effettua quindi una stima del nuovo modello (Bland, 2009).
Procedendo con l'eliminazione delle variabili non significative, si ottiene il modello ristretto `mod.coxR`, sotto riportato, in cui la significatività dei tre predittori considerati è confermata dai *p-values*.

```
> summary(mod.coxR)
Call:
coxph(formula = Surv(survival, status) ~ site + cyto5.6 + zeb1)
  n= 76
               coef exp(coef) se(coef)      z Pr(>|z|)
sitePleural -0.6182    0.5389   0.3163 -1.955   0.0506 .
cyto5.6     -0.4284    0.6516   0.1988 -2.155   0.0312 *
zeb1         0.3680    1.4449   0.1509  2.439   0.0147 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
            exp(coef) exp(-coef) lower .95 upper .95
sitePleural    0.5389     1.8556    0.2899     1.002
cyto5.6        0.6516     1.5348    0.4413     0.962
zeb1           1.4449     0.6921    1.0750     1.942

Rsquare= 0.165   (max possible= 0.996)
Likelihood ratio test= 13.71  on 3 df,   p=0.003328
Wald test            = 12.75  on 3 df,   p=0.005218
Score (logrank) test = 12.89  on 3 df,   p=0.004883
```

Appartenere al gruppo con tumori classificati come `pleural` ha l'effetto di ridurre, per `cyto 5.6`, il rischio di "decesso" di un fattore 0.652, cioè del 100 x (1 − 0.652) = 34.8%, rispetto a un paziente ma con un tumore di gruppo `peritoneal`.
Trattandosi di modelli annidati, si può effettuare il test comparativo utilizzando la funzione `anova` per verificare l'ipotesi nulla che valga il modello ridotto, contro l'alternativa che valga il modello completo, avente come esplicative tutte le variabili.

Con l'opzione `test = "Chisq"`, vengono restituiti i risultati del test del log-rapporto di verosimiglianza tra i due modelli.

La differenza tra le due devianze è di 7.08, e corrisponde al valore della statistica test log-rapporto verosimiglianza per verificare l'adattamento del modello ridotto `mod.coxR`, avente solo come variabili esplicative `site`, `cyto5.6` e `zeb1` contro il modello completo `mod.cox`. Ottenendo un livello di significatività osservato pari a 0.90 (cfr. `P(>|Chi|)`), si accetta l'ipotesi nulla, a favore del modello ridotto.

3.4.1 Il modello di Cox stratificato

Il modello di Cox prevede anche l'estensione che include gli azzardi di base diversi per sottogruppi distinti di soggetti, individuati da un certo fattore. Questo (Bland, 2009):

- consente di stratificare rispetto a variabili che non soddisfano l'assunzione di proporzionalità degli azzardi;
- consente alla funzione d'azzardo di base di variare nei diversi livelli identificati dalla variabile di stratificazione;
- è utile quando in fase di disegno dello studio abbiamo la possibilità di controllare i valori della covariata.

Supposto che x sia un fattore a S livelli, il modello di Cox stratificato è espresso come

$$\lambda(t; x_i) = \lambda_{0s}(t) \exp(\beta^T x_i),$$

con $\lambda_{0s}(t)$ funzione specifica per i soggetti nella categoria s $(s=1,\ldots, S)$ di x.

La stratificazione è fatta sulla variabile sospetta di non verificare l'assunto di proporzionalità degli azzardi, mentre nel regressore rimangono le altre covariate.

Dal modello di Cox "ristretto" si vuole quindi stratificare rispetto alla variabile dicotomica `site`, assumendo che i due gruppi producono scostamenti dalla situazione di proporzionalità. Si ottiene così il seguente modello di Cox stratificato:

```
summary(cox.strat)
Call:
coxph(formula=Surv(survival,status)~ strata(site)+cyto5.6+zeb1)
      n= 76
           coef exp(coef) se(coef)      z Pr(>|z|)
cyto5.6 -0.4257    0.6533   0.1978 -2.152   0.0314 *
zeb1     0.3114    1.3653   0.1500  2.076   0.0379 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
cyto5.6    0.6533     1.5306    0.4434    0.9628
zeb1       1.3653     0.7324    1.0175    1.8320

Rsquare= 0.128   (max possible= 0.991 )

Likelihood ratio test= 10.41  on 2 df,  p=0.005477
Wald test            = 9.55  on 2 df,   p=0.008417
Score (logrank) test = 9.78  on 2 df,   p=0.007507
```

Per le due variabili `cyto5.6` e `zeb1` si hanno *p-values* non significativi (p-values ≈ 0.03): si accetta così l'ipotesi che esse siano diverse da zero con un livello di significatività pari all'1%.

Per il modello *`cox.strat`* sopra stimato si rifiuta l'ipotesi nulla che i coefficienti siano complessivamente uguali a zero.

Quindi, per verificare l'assunto di proporzionalità del modello di Cox per la covariata `site`, si rappresenta graficamente il logaritmo degli azzardi di base cumulati rispetto al tempo: le curve risultanti dovranno risultare parallele, ovvero mantenere una distanza verticale pressoché costante. Infatti, se due azzardi sono proporzionali rispetto a una variabile x_k, il loro rapporto è costante nel tempo.

Dalla Figura 3.4, si nota che le due curve parallele, che rappresentano i due strati `peritoneal` e `pleural`, indicano che non è necessario stratificare per la variabile `site`.

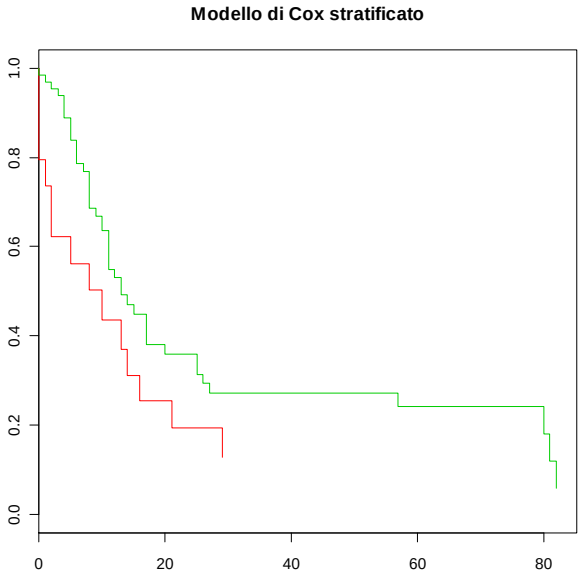

Figura 3.4: Curve di sopravvivenza per il modello di Cox stratificato rispetto a `site`.

Nella Figura 3.5, le due curve relative ai due strati `peritoneal` non si intersecano confermando che non è necessario stratificare.

Il modello di Cox stratificato non viene quindi considerato in quanto si perderebbe la possibilità di valutare le significatività dell'esplicativa `site`.

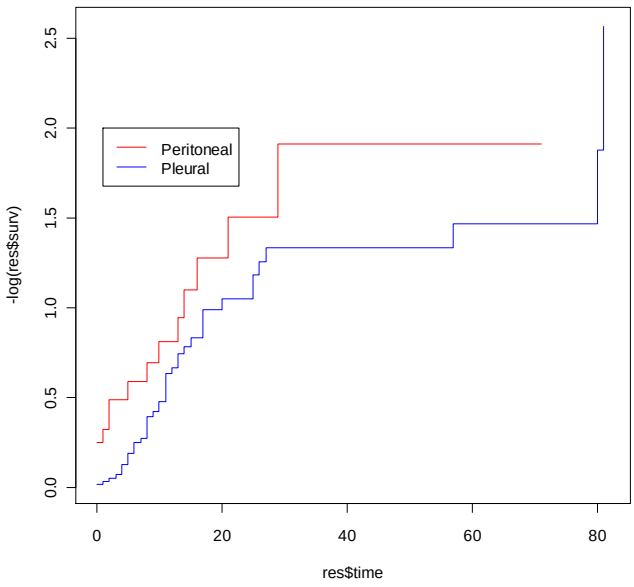

Figura 3.5: Curve relative ai due strati `Peritoneal` e `Pleural`.

3.4.2 Diagnostica nel modello di Cox

Per definizione (Klein e Moeschberger, 2003), il modello a rischi proporzionali ipotizza che i coefficienti non dipendano dal tempo *t*. Quindi prima di dichiarare valido il modello ridotto, tale ipotesi deve essere verificata.

In *R* è disponibile la funzione *cox.zph* che testa questa ipotesi per ognuna delle variabili del modello e per il modello nel suo complesso.

Si ottiene così la conferma che, l'ipotesi di rischio proporzionale su cui si basa il modello `mod.cox` non può essere respinta: infatti, come si può vedere dall'output sotto riportato, nel complesso viene stimato un `p-value GLOBAL = 0.07`.

```
> cox.zph(mod.coxR)
                rho chisq      p
sitePleural  0.224  3.43 0.0642
cyto5.6      0.206  2.28 0.1314
zeb1        -0.186  2.70 0.1003
GLOBAL          NA  7.17 0.0666
```

Inoltre, l'analisi dei residui fornisce molte informazioni riguardo l'adeguatezza delle ipotesi su cui il modello si basa e, attraverso tale analisi, è possibile individuare sia problemi riguardanti l'adattamento globale del modello ai dati, sia scostamenti isolati di singole osservazioni dal loro valore previsto (Baccini e Mealli, 2001).

Per il modello di Cox sono stati proposti diverse definizioni di residui:

1. residui di martingala;
2. residui di devianza;
3. residui parziali di *Schoenfeld*;
4. residui a *score*;
5. residui "beta".

Si presentano quindi di seguito le caratteristiche delle varie stime di residui, in riferimento al modello di Cox ridotto (`mod.coxR`).

I residui di martingala per il soggetto *i*-esimo al tempo *t* sono definiti come la differenza tra il valore $\delta_i(t)$, che indica l'appartenenza del soggetto all'insieme di rischio nel tempo *t*, e il valore della funzione di rischio cumulativa *Λ(t)* (si veda §3.1) ovvero $r_i = \delta_i(t) - \Lambda(t)$. Usualmente tale valore viene valutato al tempo in cui il soggetto esce

dall'insieme di rischio. L'impiego dei residui di martingala è suggerito da Fleming e Harrington (1991) per valutare la forma funzionale del modello a rischi proporzionali. Per il modello `mod.coxR`, come si può notare dalla Figura 3.6, vi è un buon adattamento dei residui alla retta di riferimento.

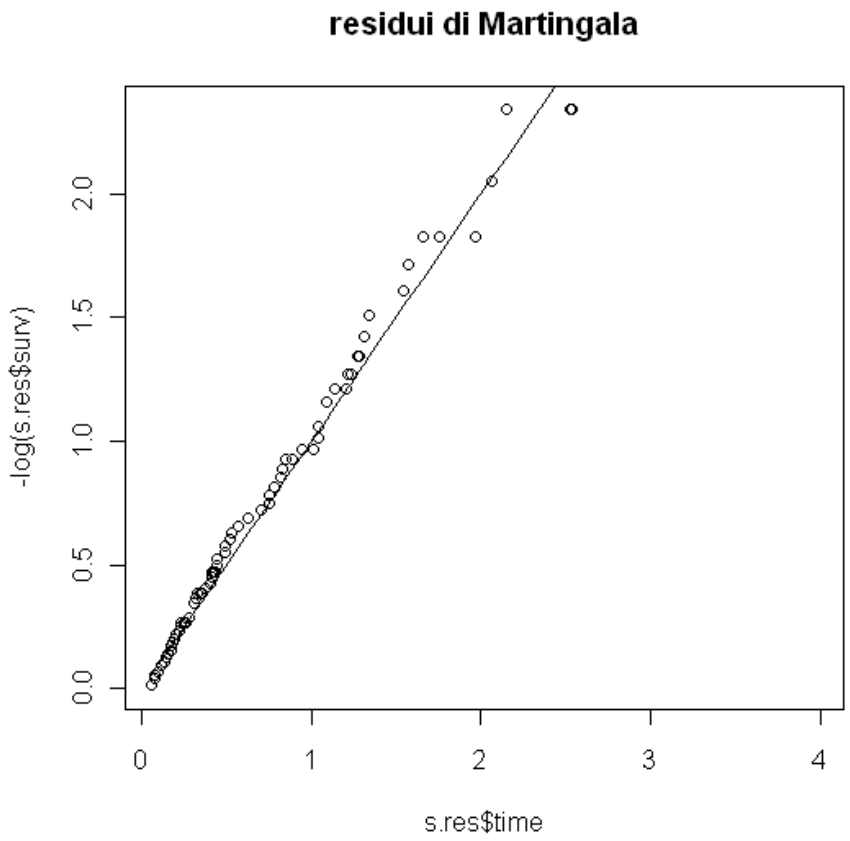

Figura 3.6: Residui di Martingala per il Modello di Cox ridotto.

Un modo semplice e frequentemente proposto in letteratura (Baccini e Mealli, 2001) per risolvere il problema dell'asimmetria dei residui appena calcolati, consiste nel considerare i residui di devianza. I residui di devianza si ottengono dai residui di martingala con una trasformazione che intende ridurre l'asimmetria, ossia $r_i^d = sign(r_i)\sqrt{-r_i - \delta_i(t)\log(\delta_i(t) - r_i)}$. Dal grafico di dispersione in Figura 3.7 (a sinistra) si osserva che i residui di devianza si dispongono in modo casuale, indice che il modello si adatta bene ai dati. Dal diagramma quantile-quantile (Figura 3.7 a destra) si

può notare che alcuni residui si discostano dalla retta di riferimento in prossimità delle code.

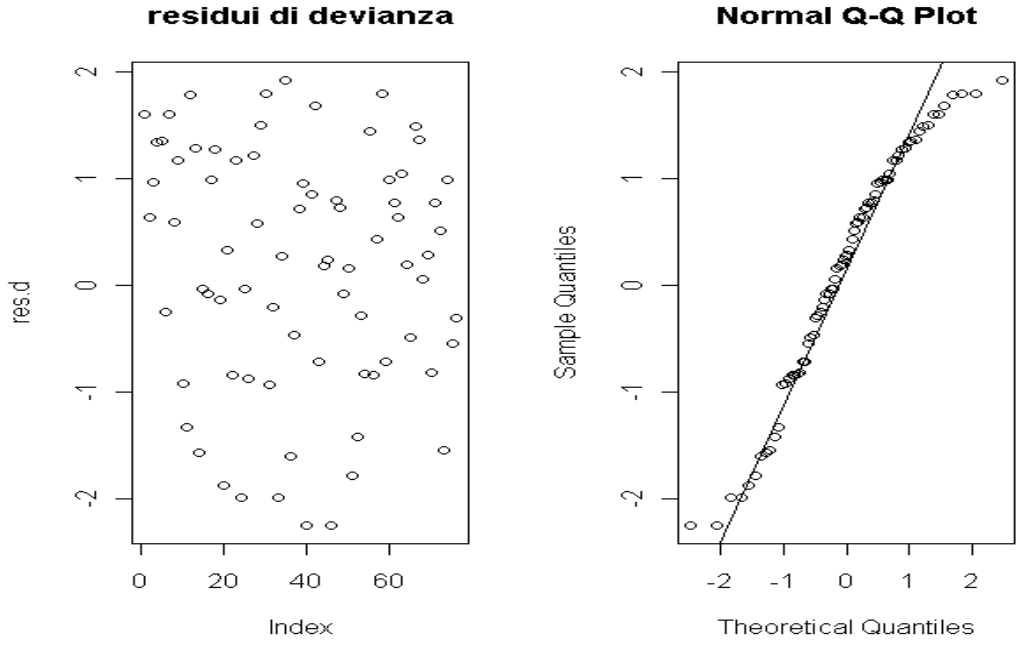

Figura 3.7: Residui di devianza per il modello di Cox ridotto.

Si possono calcolare i residui di *Schoenfeld* (Klein e Moeschberger, 2003). Essi sono calcolati per ognuna delle *k* covariate di interesse e sono definiti come il valore della covariata x_{ik} per il soggetto *i*-esimo ancora a rischio al tempo t_i meno la media ponderata della covariata, per la probabilità di fallimento $\hat{p}_j$ per ogni individuo a rischio nell'istante *j*. Sono espressi in formula come

$$r_{ik} = x_{ik} - \sum_{i=1}^{j \in R(t_i)} x_{jk}\, \hat{p}_j .$$

La Figura 3.8 illustra i residui di *Schoenfeld* calcolati per le covariate `sitePleural`, `cyto5.6` e `zeb1`. Poiché non si notano sistematiche distanze dei punti dall'area compresa tra le due linee tratteggiate, e poiché i punti si dispongono in modo casuale lungo l'asse delle ascisse (tempo) si può affermare che i residui di *Schoenfeld* sono indipendenti dal tempo. Questa loro caratteristica fondamentale (Schoenfeld,1982) porta quindi a confermare che l'assunto di proporzionalità del rischio è verificato per tutte e tre le covariate. Inoltre, non si osserva un eccesso di residui positivi o negativi in qualche regione dell'asse dei tempi e quindi sembrerebbe superfluo stratificare.

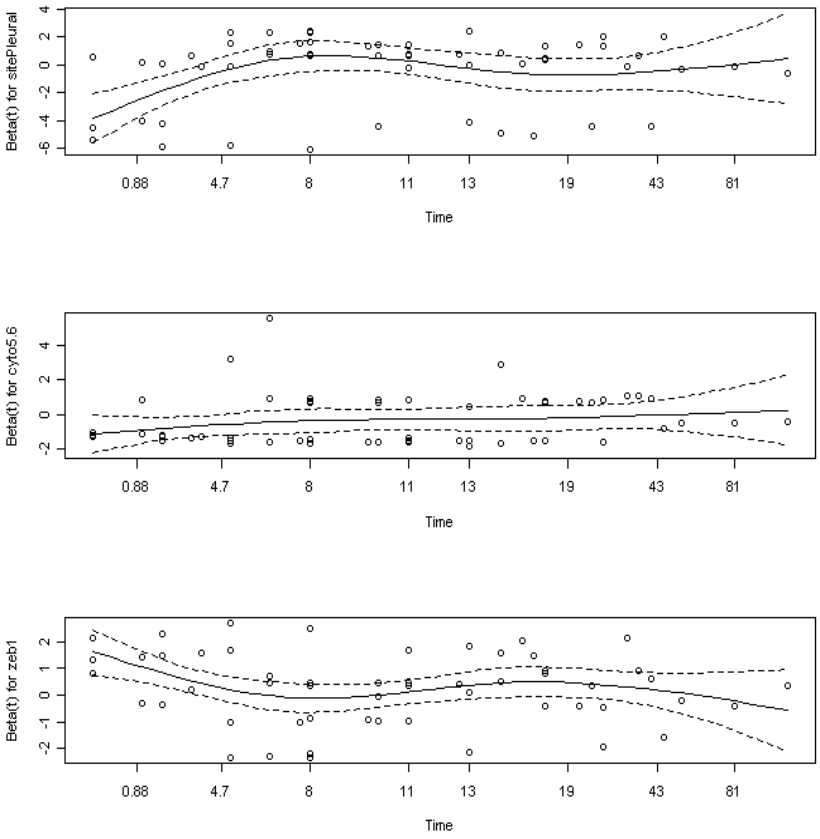

Figura 3.8: Residui di Schoenfeld distinti per covariata.

Nei modelli di regressione per dati di durata, così come nella regressione lineare, l'influenza dell'*i*-esima osservazione sulla stima del modello è espressa dal vettore (si veda ad esempio Storer e Crowley, 1985) $\Delta\beta_i = \hat{\beta}_{(-i)} - \hat{\beta}$, che costituisce la variazione indotta nella stima dei coefficienti dalla rimozione dell'osservazione *i* dal campione.

Poiché il calcolo diretto di queste differenze richiede *n* stime supplementari del modello, in letteratura sono stati proposti dei metodi per approssimare i $\Delta\ \beta_i$. Uno di questi è basato sui residui *score*, che costituiscono una matrice di *n* righe e colonne pari al numero di predittori, e che vengono impiegati per valutare se una osservazione è influente nel calcolo dei valori stimati dei coefficienti di regressione.

I residui *score*, riportati in Figura 3.9, evidenziano il contributo dei singoli soggetti alla stima del modello, con degli *outlier*s corrispondenti:

- per `sitePleural` alle osservazioni 46, 20, 26 e 52;

- per `cyto5.6` all'osservazione 42;
- per `zeb1` alle osservazioni 20, 46, 52 e 40.

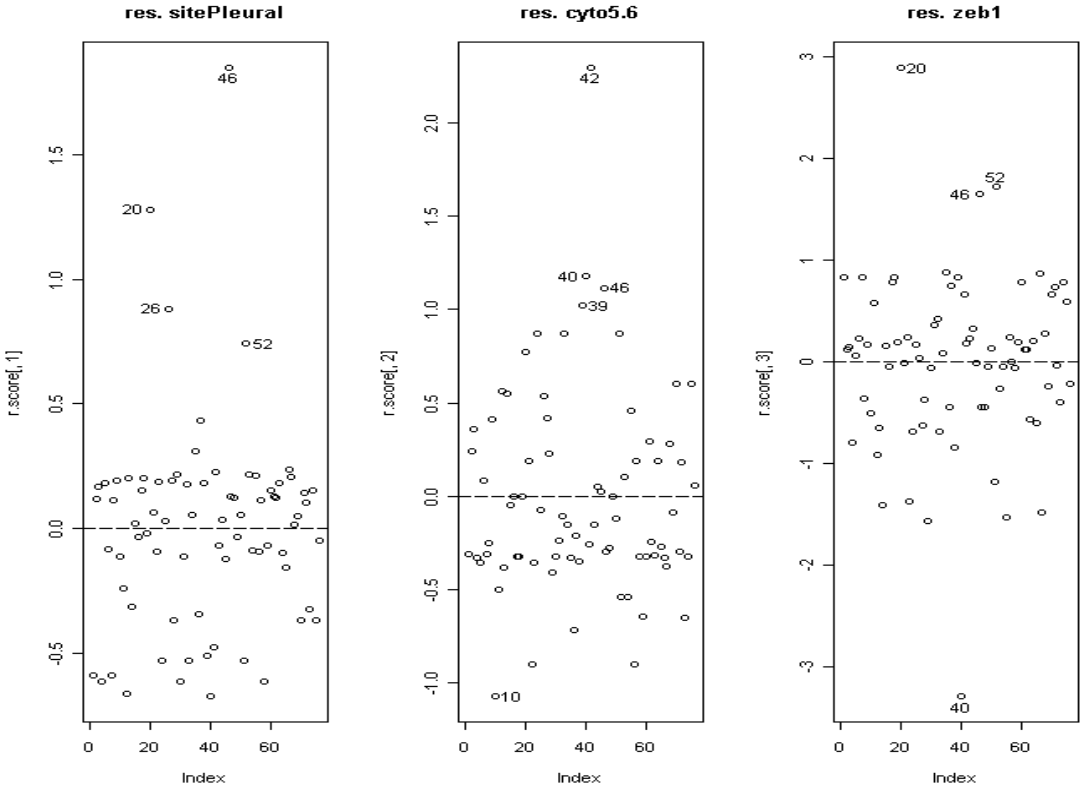

Figura 3.9: Residui *score* per le variabili del modello di Cox ridotto.

Infine utilizzando i residui beta (standardizzati) è possibile verificare la presenza di punti leva per ciascuna variabile stimata dal modello. I punti leva hanno maggior peso nel determinare l'andamento della retta di regressione e non dovrebbero superare il valore critico dato da $\frac{2 \ x \ p}{n}$, con *p* numero di regressori, ovvero nel caso in esame $\frac{2 \ x \ 3}{76} = 0.08$.

Dalla Figura 3.10, in cui sono rappresentati i residui *Beta*, risultano evidenti solo 4 residui dell'osservazione 46 per `sitePleural`, della 42 per `cyto5.6` e delle osservazioni 20 e 40 per `zeb1`, che corrispondono alle osservazioni individuate nei residui *score*.

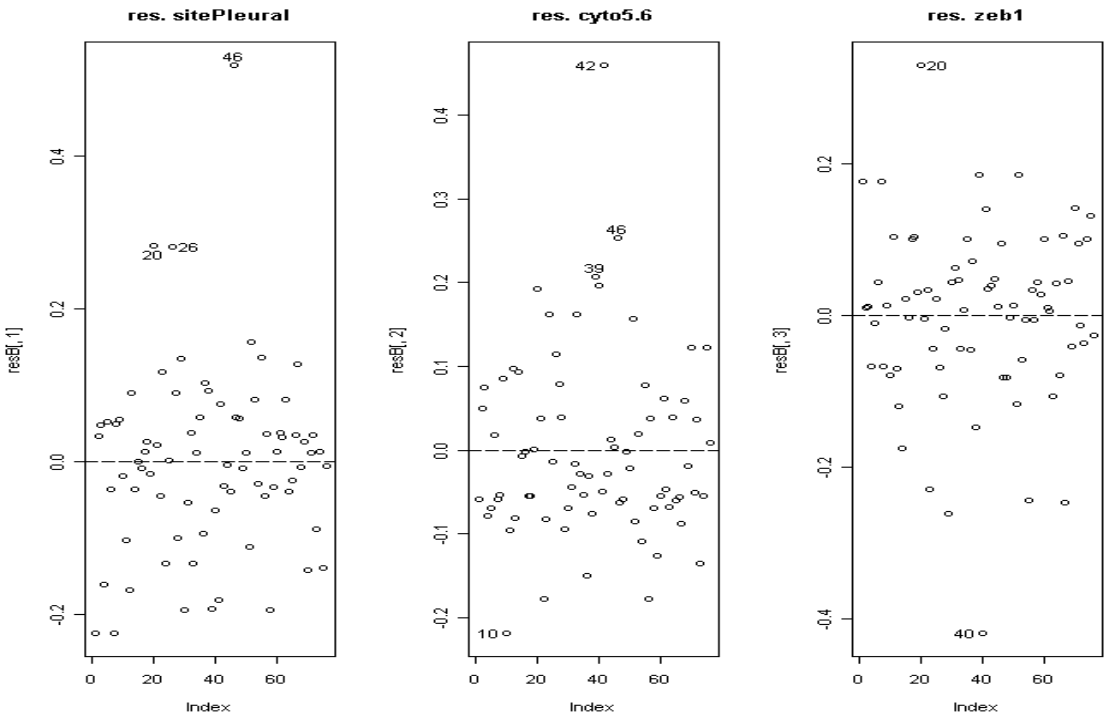

Figura 3.10: Residui Beta per le variabili del modello di Cox ridotto.

Dai *Q-Q plot* riportati in Figura 3.11 si può verificare che i residui *Beta* sono distribuiti in modo normale anche se, sulle code, si notano degli allontanamenti dalla retta di riferimento.

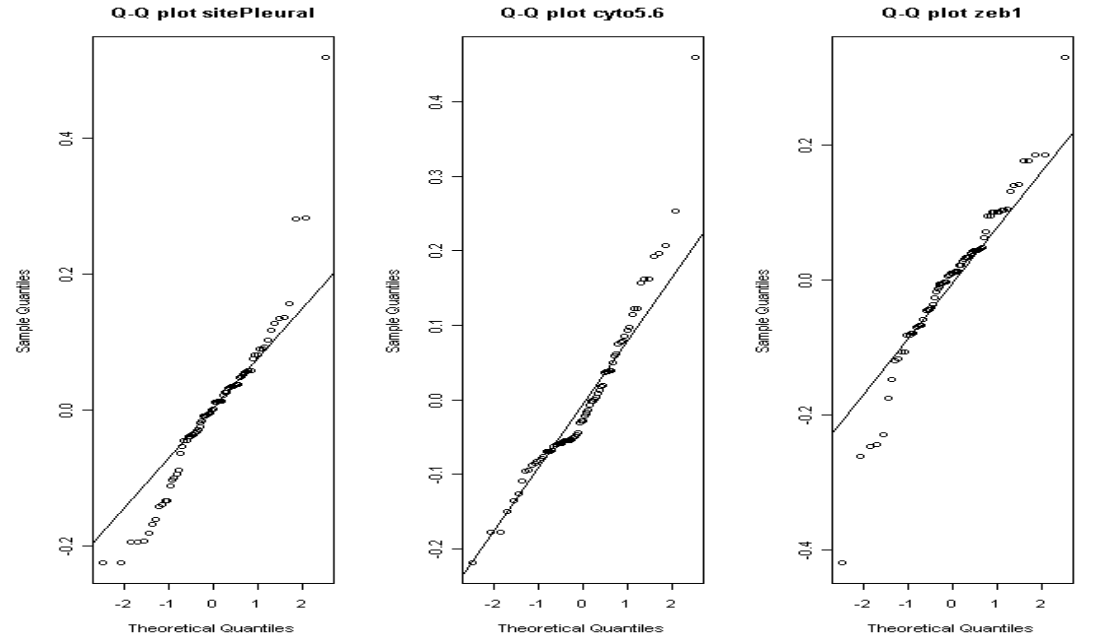

Figura 3.11: *Q-Q plot* dei residui "Beta" per le variabili del modello di Cox ridotto.

3.5 Considerazioni finali

Dall'analisi dei residui calcolati per il modello di Cox ridotto, in cui risultano significative solo le tre variabili `site`, `cyto5.6` e `zeb1`, si può concludere che, complessivamente, il modello `mod.coxR` è discreto nella spiegazione del dataset.

Non può essere però tralasciata la presenza di quattro osservazioni anomale (20, 40, 42, 46) evidenziate sia nel grafico dei residui *beta* che in quello dei residui *score*.

Nel prossimo capitolo si propone quindi un modello robusto rispetto agli *outliers* o a eventuali specificazioni scorrette, stimato con i "*Robustly Proportional Hazards*". Quest'alternativa robusta al *PLE*, è basata su una doppia stima parziale della verosimiglianza che modifica la stima dell'equazione $\sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j \geq i} \exp(x_j^T \beta) x_j}{\sum_{j \geq i} \exp(x_j^T \beta)} \right] = 0,$

senza fondamentalmente cambiare la struttura del modello (Heritier *et al*., 2009).

CAPITOLO 4

ANALISI DELLA SOPRAVVIVENZA ROBUSTA

4.1 Introduzione

E' opportuno, prima di entrare nel vivo dell'analisi della sopravvivenza robusta, effettuare una premessa fondamentale alla comprensione di alcune nozioni che verranno menzionate nei paragrafi successivi.

La statistica parametrica classica, in particolare, a seconda delle informazioni disponibili, si basa sull'assunzione di un modello statistico parametrico specificato da

$$\Im = \{ f(y;\theta), \theta \in \Theta \},$$

dove $f(y;\theta)$ denota una funzione di densità, con $\Theta \subseteq R^p$ ($p \geq 1$) spazio parametrico.

Il modello $\Im$ copre un ruolo centrale nelle procedure classiche di inferenza. Infatti, quando si specifica un modello statistico $\Im$ per i dati, si ipotizza che, idealmente, il modello probabilistico generatore dei dati appartenga a $\Im$, ossia che il modello sia correttamente specificato (si veda ad esempio Pace e Salvan, 2001, Cap. 12).

Talvolta, però il modello $\Im$ può non rispecchiare esattamente la realtà, o per la presenza di dati anomali (*outliers*) nel campione osservato o per il carattere approssimato del modello teorico stesso, non rispondendo così a una sorta di principio di stabilità: piccoli spostamenti dal modello ipotizzato non dovrebbero produrre grossi cambiamenti inferenziali. Ciò è particolarmente vero in ambito parametrico, in quanto diverse procedure classiche risultano molto, ed a volte estremamente, sensibili a piccoli spostamenti della distribuzione dei dati dal modello assunto.

Quando una procedura inferenziale risulta sensibile rispetto a piccole deviazioni, sono allora necessari metodi statistici affidabili in un qualche intorno del modello. Una procedura statistica (nella pratica uno stimatore, una statistica test, etc.) poco sensibile a piccoli o moderati scostamenti (in termini della distribuzione dei dati) dal modello ipotizzato è detta *ROBUSTA*. La teoria della robustezza descrive allora le proprietà delle procedure statistiche in un intorno del modello parametrico. Essa costituisce un approccio alla statistica, non un ramo della statistica, in quanto ha lo scopo di salvaguardare rispetto a eventuali deviazioni dalle ipotesi statistiche assunte. Alcuni

riferimenti utili sono Hampel *et al.* (1986), Huber e Ronchetti (2009) e Heritier *et al.* (2009).

4.2 La robustezza

Nello studio di un modello statistico sono possibili vari approcci per l'analisi dei dati. La Figura 4.1 raffigura tre metodi per analizzare i dati (cfr. Hampel *et al.*,1986):

a) procedure inferenziali classiche (es. stima dei minimi quadrati);
b) metodi basati sull'eliminazione delle osservazioni anomale;
c) metodi robusti con buone proprietà in termini di efficienza.

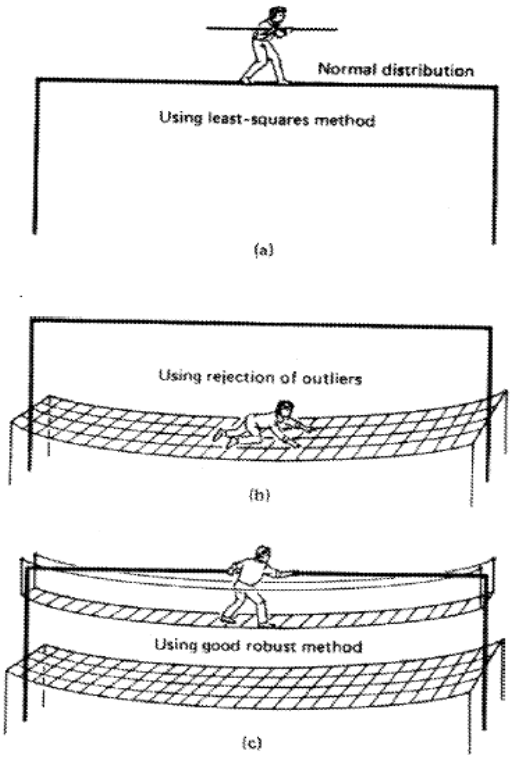

Figura 4.1: Vari metodi per analizzare i dati

Le procedure parametriche classiche implicano l'efficienza, mentre quelle basate sull'eliminazione delle osservazioni anomale la stabilità. La statistica robusta si colloca pertanto in una situazione intermedia.

La principale critica mossa alla robustezza è che qualunque statistico accorto si muoverebbe adottando una qualche procedura di individuazione e eliminazione delle osservazioni anomale, procedendo poi, solo in un secondo momento, con l'applicazione delle procedure classiche adeguate per l'inferenza.

Obiettivo della statistica robusta è predisporre strumenti per valutare la bontà delle procedure statistiche in intorni di modelli, e quindi di trovare procedure che mantengano buone proprietà anche quando il modello ipotizzato è solo un'approssimazione del "vero" modello (Pelegatti, 2000).

La teoria della robustezza:

- descrive le procedure di inferenza necessarie a prevenire perdite di efficienza e consistenza rispetto a eventuali deviazioni dal modello;
- costituisce un approccio alla statistica in quanto ha lo scopo di prevenire eventuali effetti causati da valori anomali o da osservazioni influenti.

Si possono distinguere due aspetti di robustezza:

1) rispetto alla contaminazione (*outliers*): la presenza di dati anomali può essere dovuta a errori di rilevazione o di codificazione. Anche una lieve eterogeneità della popolazione, riconducibile a distribuzioni con code pesanti e imputabile a una definizione del modello statistico non del tutto accurata, può produrre dati anomali (Pace e Salvan, 2001, Cap. 12);
2) rispetto alla specificazione scorretta *(misspecification)*: il modello statistico, pur catturando qualitativamente e quantitativamente aspetti importanti dei dati, potrebbe non descrivere con esattezza tutti gli aspetti della variabilità della popolazione a causa del uso carattere approssimato.

Talvolta, dal punto di vista pratico, possono essere equivalenti. Infatti, una scorretta specificazione del modello può essere causa della presenza di dati anomali, ossia di osservazioni distanti dalla maggioranza dei dati.

4.2.1 Equazioni di stima non distorte

Sia $F(y;\theta)$ la funzione di ripartizione associata a $f(y;\theta)$. Un modello che tiene conto di una possibile osservazione anomala nel punto x, che si verifica con probabilità ε, ha funzione di ripartizione

$$F\left(y;\theta,\varepsilon,x\right)=(1-\varepsilon)F\left(y;\theta\right)+\varepsilon\delta_x,$$

e rappresenta una contaminazione della distribuzione $F\left(y;\theta\right)$ con una distribuzione degenere in x, dove δ_x indica la funzione di ripartizione di una variabile casuale degenere nel punto x.

Con x fissato e quando la contaminazione è infinitesimale, ovvero per $\varepsilon\rightarrow 0$, si desidera applicare procedure di inferenza su θ che siano robuste, ovvero che non risentano

della scorretta specificazione di $\Im$. Per un modello con verosimiglianza regolare (ovvero derivabile due volte) e sotto ulteriori assunzioni, di cui la principale è che il supporto di Y non dipenda da θ, si ha che $E_\theta(l_*(\theta)) = 0$ per ogni $\theta \in \Theta$; ovvero l'equazione di verosimiglianza $l_*(\theta) = 0$ è un'equazione di stima non distorta (Pace e Salvan, 2001, Cap. 12). Lo stimatore di massima verosimiglianza $\hat{\theta}$, ottenuto sotto $\Im$, rimane allora consistente.

In generale, un'equazione di stima, è un'equazione in θ e in y del tipo $q(\theta; y) = 0$ che, risolta, fornisce una stima $\tilde{\theta}$ di θ. La funzione $q(\theta; y)$ è detta quindi funzione di stima.

Nell'ambito del modello $\Im$ con parametro $\theta \in \Theta \subseteq \Re^p$, l'equazione di stima $q(\theta; y) = 0$ è detta non distorta se $E_\theta(q(\theta, Y)) = 0$, per ogni $\theta \in \Theta$.

In corrispondenza di un campione casuale semplice di numerosità n, usualmente, si hanno funzioni di stima della forma

$$q(\theta; y) = \sum_{i=1}^{n} g(\theta; y_i), \tag{4.1}$$

per una data funzione $g(\cdot)$.

4.2.2 La funzione d'inferenza

L'approccio infinitesimale è basato su tre concetti fondamentali: robustezza qualitativa, funzione d'influenza e punto di rottura.

Una statistica T è qualitativamente robusta se moderate deviazioni della distribuzione ipotizzata comportano solamente moderate deviazioni della legge di T, per ogni ampiezza campionaria.

La funzione di stima non distorta (4.1) può essere scritta come $\frac{1}{n}\sum_{i=1}^{n} g(\theta; y_i)$, ossia come

$$E_{\hat{F}_n}(g(\theta; Y_i)), \tag{4.2}$$

dove $E_{\hat{F}_n}$ indica il valore atteso rispetto alla distribuzione empirica con funzione di ripartizione $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, x]}(y_i), x \in \Re$.

Dunque si può pensare di rappresentare $\tilde{\theta}_n$, stima di θ ottenuta dall'equazione di stima corrispondente alla (4.2), come una funzione di $\hat{F}_n$, ossia $\tilde{\theta}_n = T(\hat{F}_n)$. Poiché $\hat{F}_n$ è una stima della funzione di ripartizione $F = F(y;\theta)$, il parametro oggetto di inferenza

può essere indicato come $\theta = T(F)$. La funzione *T(·)* dallo spazio delle funzioni di ripartizione univariate in Θ è detta funzionale statistico.

Lo strumento più noto per l'approccio infinitesimale è la funzione d'influenza, introdotta nel 1974 da Hampel, inizialmente col nome di curva d'influenza. Essa descrive l'effetto di una contaminazione infinitesimale nel punto *x*, effetto standardizzato secondo la massa ε della contaminazione. Formalmente, la funzione d'influenza *IF(x;T,F)* per *T* è definita come

$$IF(x;\, T,\, F) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon}.$$

Vi sono alcune quantità derivate dalla *IF* che ne riassumono gli aspetti rilevanti per quanto riguarda lo studio della robustezza di una statistica. La più importante è la *gross-error sensitivity* che misura la sensibilità di una statistica alla presenza di *outliers*, definita da

$$\gamma^* = \sup_x |IF(x;\, T,\, F)|.$$

Ad essa è legata la nozione di *B-robustezza* (dove *B* sta per *bias*), che può infatti essere vista come robustezza rispetto agli *outliers*. La statistica *T* viene definita *B-robusta* sotto la distribuzione *F* se $\gamma^* < +\infty$. Quindi, lo stimatore $\tilde{\theta}$ di θ *è* robusto, nel senso di avere indice γ* limitato se, in generale, attribuisce peso contenuto alle osservazioni anomale (Pace e Salvan, 2001, Cap. 12).

Altra misura di robustezza derivata dalla *IF* è la *local-shift sensitivity*, ovvero la sensibilità rispetto a fluttuazioni locali di *T* sotto la distribuzione *F*, data da

$$\lambda^* = \sup_{x \neq y} \frac{|IF(y;T,F) - IF(x;T,F)|}{|y - x|}.$$

Questa quantità rappresenta l'effetto dello spostamento di un'osservazione da *x* in *y*. Se $\lambda^* < +\infty$ si ha la robustezza di *T* rispetto ad errori di arrotondamento.

Come terza misura di robustezza, sempre derivata dalla *IF*, si ha il punto di rifiuto (*rejection point*) dello stimatore *T* sotto la distribuzione *F*, definito da

$$\rho^* = \inf \{ r > 0 : IF(x, T, F) = 0 \; per \; |y| > r \},$$

con $\rho^* = \infty$ se non esiste un tale *r*. Tale misura rappresenta il punto di rigetto dei valori anomali da parte di *T*. Infatti, se $\rho^* < \infty$, la contaminazione provocata da punti *y* con $|y| > \rho^*$ non esercita alcuna influenza sul valore di *T*.

Infine, dato che la funzione d'influenza consente uno studio solamente locale del comportamento di una statistica, è necessario affiancarle un uno strumento che misuri

la robustezza in modo globale: il punto di rottura assolve proprio questo compito. Informalmente, esso può essere definito come la più piccola frazione di osservazioni (anomale) che rendono il funzionale inaffidabile. Il punto di rottura (*breakdown point*) della statistica *T* rappresenta la percentuale massima di osservazioni anomale che la statistica *T* può sopportare (cfr. Hampel *et al*., 1986, Cap.2).

4.2.3 Stimatori di tipo M

Una scorretta specificazione del modello statistico, come ad esempio uno scostamento più o meno marcato dall'ipotesi di normalità, può portare, a seconda dei casi, a pesanti conseguenze sulle procedure inferenziali. Pertanto, nelle situazioni in cui non abbiamo sufficienti informazioni sul fenomeno d'interesse, provenienti dai dati o da altre fonti, è auspicabile ricorrere a statistiche robuste rispetto alla specificazione scorretta.

Un'importante classe di stimatori legata a funzioni di stima robuste, introdotta da Huber nel 1964, è quella degli stimatori di tipo M, ovvero "*maximum likelihood type estimators*" in quanto sono una generalizzazione degli stimatori di massima verosimiglianza (Pelegatti, 2000).

Si definisce, stimatore di tipo M per θ uno stimatore che soddisfa l'equazione

$$g(\theta; y) = \sum_{i=1}^{n} g(\theta; y_i) = 0, \tag{4.3}$$

dove g $(\cdot)$ è una funzione nota con valori in $\Re^p$.

Sotto opportune condizioni di regolarità (Bernarski, 1993), si può dimostrare che lo stimatore $\tilde{\theta}$ è consistente e con distribuzione asintotica normale, ossia

$$\tilde{\theta} \sim N_p(\theta, V(\theta)),$$

dove $V(\theta)$ è una matrice data da

$$V(\theta) = M(\theta)^{-1} Q(\theta) M(\theta)^{-T},$$

con $M(\theta) = -E\left[\frac{\partial}{\partial \theta} g(\theta; y)\right]$ e $Q(\theta) = E\left[g(\theta; y)\, g(\theta; y)^T\right]$ (Hampel *et al.*, 1986).

Si dimostra (cfr. Hampel *et al.*, 1986) che, per gli stimatori di tipo M della forma (4.3), la funzione d'influenza è

$$IF(y; T, F) = M(\theta)^{-1} g(\theta; y)$$

Pertanto, l'indice γ^* è finito se e solo se la funzione di stima $g(\theta; y)$ è limitata.

- *<u>Esempio</u>: Modello di posizione e stimatore di Huber*

Per un modello di posizione, ossia tale che F*(y;θ) = F₀(y-θ)*, è naturale scegliere la funzione $g(\theta; y)$ della forma $g(\theta; y) = y - \theta$. In tale situazione, l'espressione della *IF* e della varianza asintotica di uno stimatore di tipo M per parametro di posizione si riducono a (Hampel *et al.*,1986)

$$IF(x; \tilde{\theta}) = \frac{g(x)}{\int g'(y) dF_0(y)}$$

e

$$V(\theta) = \frac{\int g(y)^2 dF_0(y)}{\left[\int g'(y) dF_0(y)\right]^2}.$$

Si noti anche che, se *F* è assolutamente continua, tali espressioni per la *IF*, e la varianza asintotica mantengono la loro validità anche se la funzione *g (·)* è derivabile a meno di un numero finito di punti.

Nel caso particolare in cui le $y_1, \ldots, y_n$ sono osservazioni indipendenti tratte da una normale *N(θ,1)*, la stima di massima verosimiglianza $\bar{y}$ è basata su una funzione di stima del tipo $q(\theta; y) = \sum_{i=1}^{n} g(\theta; y_i)$, con $g(\theta; y_i) = y_i - \theta$. Si tenga presente che, essendo proporzionale a *c-θ*, la corrispondente funzione d'influenza non è limitata. Uno stimatore robusto rispetto a contaminazioni infinitesimali può allora essere definito sostituendo la funzione $g(\theta; y_i) = y_i - \theta$ con la funzione limitata $g_b(\theta; y_i) = [y_i - \theta]_{-b}^{b}$, dove *b* è una costante positiva assegnata e la notazione $[h(x)]_{-b}^{b}$, per una generica funzione *h*(*x*), indica il suo troncamento ai valori *b* e *-b*, ossia

$$[h(x)]_{-b}^{b} = \begin{cases} -b & se \quad h(x) < -b \\ h(x) & se \quad -b \le h(x) \le b \\ b & se \quad h(x) > b \end{cases}$$

Lo stimatore di Huber, introdotto nel 1964, è quindi quel valore di θ per cui vale $\sum_{i=1}^{n} [y_i - \theta]_{-b}^{b} = 0$. Esso gode di numerose proprietà interessanti. Ad esempio, esso ha varianza asintotica minima nella classe degli stimatori Fisher consistenti con γ^* limitato (Hampel *et al.*,1986). Il valore di *b* va scelto in modo tale da raggiungere un

compromesso soddisfacente tra robustezza ed efficienza. Una scelta comune per *b* è il valore 1.345. Quando *b* tende a 0, lo stimatore di Huber tende alla mediana, che è lo stimatore di tipo M più *B-robusto* ma meno efficiente.

Si tenga conto che (Pelegatti, 2000) lo stimatore di Huber ha il difetto di non essere invariante a trasformazioni di scala. Per renderlo tale si può usare $\sum_{i=1}^{n} \left[\frac{y_i - \theta}{s} \right]_{-b}^{b} = 0$, dove *s* è una stima robusta del parametro di scala e può essere calcolata per mezzo della mediana delle deviazioni assolute dalla mediana, detta MAD (*Median Absolute Deviation*), data da

$$MAD\ (y) = med\ \left(\left\{ \left| y_i - med\ \left(\{ y_i \}_{i=1,\ldots,n} \right) \right| \right\}_{i=1,\ldots,n} \right)$$

A scopo illustrativo, nella Tabella 4.1 si riporta lo stimatore di tipo M di Huber per il parametro di posizione delle varie variabili esplicative del *dataset*, con $b = 1.345$ e con scala stimata con il *MAD*.

	survival	eta	ecad	ncad	bcad	mmp2	mmp9	cyto5.6	vim	SMA	zeb1	zeb2	S100A
MAD	8.89	8.89	1.48	1.48	1.48	0.00	1.48	0.00	1.48	0.00	1.48	0.00	1.48
$[h(x)]_{-b}^{b}$	13.13	69.86	0.97	1.58	1.53	n.c.	0.97	n.c.	1.97	n.c.	1.71	n.c.	1.31
mediana	11.00	70.00	1.00	2.00	2.00	0.00	1.00	0.00	2.00	0.00	2.00	0.00	1.00

Tabella 4.1: Stima di Huber calcolata per le variabili del dataset.

Si possono fare delle osservazioni sulla stima di Huber:

- l'equazione che definisce lo stimatore M ha un'unica soluzione;
- se $[h(x)]_{-b}^{b}$ non è limitata, lo stimatore non è né *B*-robusto, né qualitativamente robusto, e ha punto di rottura uguale a 0;
- se $[h(x)]_{-b}^{b}$ è limitata, lo stimatore è *B*-robusto, qualitativamente robusto, e con punto di rottura paria a 0.5.

Nel caso in esame, per le proteine `mmp2`, `cyto5.6`, `SMA` e `zeb2` il *MAD* risulta pari a zero: la stima di Huber risulta pertanto non calcolabile; in Tabella 4.1 viene quindi riportato "n.c.". In particolare, per le variabili `eta`, `ecad`, `mmp9` e `vim` la stima $[h(x)]_{-b}^{b}$ ottenuta è approssimabile alla mediana. Inoltre, si nota che il *MAD* per le due

variabili quantitative `survival` ed `eta` vale 8.89, mentre per le proteine la mediana delle deviazioni assolute dalla mediana coincide, per 7 casi su 11, al valore 1.48.

4.3 Inferenza robusta nel modello di Cox

In questo capitolo si presenta un'alternativa robusta al modello di Cox a rischi proporzionali, dove le variabili esplicative sono indipendenti dal tempo, con un evento per ogni soggetto.

Come emerso da uno studio di Bednarski (1993), il modello di regressione robusta ("*Robustly Proportional Hazards*") introduce nella massimizzazione della verosimiglianza parziale una doppia pesatura del *PLE* che modifica la stima dell'equazione $\sum_{i=1}^{n} \delta_i \left[x_i - \frac{S^{(1)}(t_i;\beta)}{S^{(0)}(t_i;\beta)} \right] = 0$, senza fondamentalmente influire sulla struttura del modello di Cox nella sua forma base (si veda § 3.3).

Siano dati dei pesi opportuni, spiegati successivamente in dettaglio (4.4), della forma $w_{ij} = w(t_i, x_j)$ e $w_i = w_{ii} = w(t_i, x_i)$, con $1 \le i \le j \le n$. Si definiscono quindi le due somme

$$S_w^{(0)}(t_i;\beta) = \sum_{j \ge i} w_{ij} \quad \exp(x_j^T \beta),$$

$$S_w^{(1)}(t_i;\beta) = \sum_{j \ge i} w_{ij} \quad \exp(x_j^T \beta)\, x_j .$$

La soluzione robusta in β è quindi ottenuta risolvendo

$$U_w = \sum_{i=1}^{n} U_{w,i} = \sum_{i=1}^{n} w_{ij} \delta_i \left[x_i - \frac{S_w^{(1)}(t_i;\beta)}{S_w^{(0)}(t_i;\beta)} \right] = 0,$$

con $U_{w,i} = w_{ij} \delta_i \left[x_i - \frac{S_w^{(1)}(t_i;\beta)}{S_w^{(0)}(t_i;\beta)} \right]$ *score* individuale che rappresenta il contributo dell'*i*-esima osservazione alla sommatoria. La funzione U_w è pertanto lo *score* totale.

Bednarski (1993), e successivamente Minder e Bednarski (1996), utilizza l'azzardo di base $\lambda_0(t)$, ma ritiene che quando il tempo t è troppo grande il modello esponenziale $\lambda o(t)\ exp(x^T \beta)$ fallisce (Heritier *et al*, 2009). Sulla base di opportuni troncamenti, ha quindi proposto di utilizzare le seguenti funzioni dei pesi

$$w(t,x)=\begin{cases} k-\min(K,\lambda_0(t)\exp(x^T\beta)) & \textit{lineari} \\ \exp(-\lambda_o(t)\exp(x^T\beta/K) & \textit{esponenzia li} \\ \max(0,K-\lambda_0(t)\exp(x^T\beta))^2/K^2 & \textit{quadratici}\ , \end{cases} \tag{4.4}$$

dove *K* è il valore di troncamento che è necessario stabilire a priori per il calcolo dei pesi, in quanto definisce il *cut-off* tra efficienza e robustezza.

Scelto un valore specifico per il quantile τ della distribuzione empirica, questo viene poi utilizzato per ricavare il valore di troncamento *K*. Ad esempio, se si sceglie τ = 90% (o 95%) significa che il 10% (o 5%) dei pesi sarà considerato pari a zero dal modello a rischi proporzionali. Ciò automaticamente genera una perdita di efficienza del modello in termini di osservazioni che verranno ignorate.

Un vantaggio di questo schema di ponderazione adattiva basata sulla stima del rischio cumulativo è che questo *adaptive robust estimator* (*ARE*) è invariante rispetto alle trasformazioni del tempo: sia $\lambda o(t)$, che il vettore dei parametri β, sono stimati iterativamente con regressione robusta.

Sotto regolari condizioni su *w(t,x)*, lo stimatore ARE è consistente e si distribuisce asintoticamente come

$$\sqrt{n}(\hat{\beta}_{ARE}-\beta)\rightsquigarrow \mathrm{N}(0,V_w(\beta)),$$

dove la varianza asintotica è ricavata dallo stimatore *sandwich*

$$V_w(\beta)=M_w^{-1}Q_wM_w^{-T}.$$

Bednarski (1996) suggerisce di calcolare le matrici $M_w=M_w(\beta)$ e $Q_w=Q_w(\beta)$ come

$$\hat{M}_w(\beta)=-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_{w,i}}{\partial\beta} \quad \text{e} \quad \hat{Q}_w=\frac{1}{n}\sum_{i=1}^{n}U^*_{w,i}U^{*T}_{w,i}\ ,$$

dove $U_{w,i}$, è lo *score* individuale e $U^*_{w,i}=U_{w,i}-C_{w,i}(\beta)$, con

$$C_{w,i}(\beta)=\exp(x_i^T\beta)x_i\sum_{j\le i}\frac{w_i\delta_j w_{ji}}{S_w^{(0)}(t_j;\beta)}-\exp(x_i^T\beta)\sum_{j\le i}\frac{w_j w_{ji}\delta_j S_w^{(1)}(t_j;\beta)}{\left[S_w^{(0)}(t_j;\beta)\right]^2}.$$

Si può dimostrare che l'*IF* empirica per lo stimatore *ARE* (Heritier *et al.*, 2009) è proporzionale allo *score shiftato*, ossia

$$\hat{IF}_{w,i}=\hat{M}_w^{-1}(\beta)U^*_{w,i}(\beta)=\hat{M}_w^{-T}(\beta)(U_{w,i}(\beta)-C_{w,i}(\beta)).$$

Il test robusto di *Wald* per la verifica d'ipotesi del tipo $H_0 : \beta_{(2)} = \beta_{(2)}^0$, con $\beta = (\beta_{(1)}, \beta_{(2)})^T$, assume la forma

$$W_w = n(\hat{\beta}_{ARE\,(2)} - \beta_{(2)}^0)^T \hat{V}_{w(22)}{}^{-1} (\hat{\beta}_{ARE\,(2)} - \beta_{(2)}^0),$$

dove $\hat{V}_{w(22)}$ è il blocco (2,2) della varianza asintotica $\hat{V}_w(\beta)$. Sotto l'ipotesi nulla, W_w si distribuisce come un χ^2 con $k = \dim(\beta_{(2)}^0)$ gradi di libertà.

4.4 Analisi dei dati

Per applicare l'alternativa robusta alla stima della verosimiglianza parziale (*PLE*) e calcolare il test di Wald robusto, menzionati nel paragrafo precedente, si utilizza la libreria `coxrobust` di *R*.

Per un confronto più immediato con i risultati ottenuti dal modello di Cox, nella prima colonna della Tabella 4.3 si riportano, le stime del modello *PLE*. Quindi nelle successive colonne vi sono le stime *ARE* dei coefficienti β ottenuti per il *dataset*, considerando le tre possibili funzioni dei pesi $w(t, x)$ in (4.4).

Come quantile τ della distribuzione empirica, utilizzato da *R* per ricavare il valore di troncamento *K*, si è ritenuto opportuno scegliere $\tau = 0.95$; così il 5% dei pesi è considerato pari a zero dal modello a rischi proporzionali.

Nel caso in esame, dalla Tabella 4.3, si direbbe non sussistano rilevanti differenze in termini di coefficienti *ARE* stimati per le tre tipologie di pesi. Anche i *p-values* ottenuti, indipendentemente dai pesi utilizzati, evidenziano che molte stime risultano non significative ai fini dell'analisi (*p-value* > 0.05).

Modello completo	PLE		ARE τ=0.95					
			Pesi lineari		Pesi esponenziali		Pesi quadratici	
variabile	coef (s.e.)	p-value	coef (s.e.)	p-value	coef (s.e.)	p-value	Coef (s.e.)	p-value
eta	-0.02 (0.02)	0.232	-0.02 (0.03)	0.578	-0.02 (0.03)	0.444	-0.01 (0.03)	0.607
sessoM	-0.02 (0.32)	0.948	-0.42 (0.37)	0.257	-0.22 (0.48)	0.650	-0.50 (0.39)	0.196
Site Pleural	-0.84 (0.42)	0.045	-0.96 (0.60)	0.105	-0.90 (0.55)	0.099	-0.95 (0.61)	0.120
Subtype Mixed	-0.41 (0.52)	0.435	-0.62 (0.68)	0.366	-0.48 (0.67)	0.472	-0.73 (0.73)	0.318
Subtype Sarcomatous	0.06 (0.58)	0.913	0.25 (0.70)	0.719	0.18 (0.68)	0.780	0.32 (0.71)	0.656
ecad	0.22 (0.19)	0.256	0.25 (0.24)	0.284	0.22 (0.23)	0.339	0.27 (0.25)	0.286
ncad	0.02 (0.23)	0.910	-0.09 (0.29)	0.748	-0.05 (0.25)	0.845	-0.18 (0.29)	0.541
bcad	0.07 (0.25)	0.769	0.24 (0.31)	0.450	0.16 (0.31)	0.615	0.28 (0.32)	0.392
mmp2	0.07 (0.22)	0.753	0.19 (0.30)	0.524	0.13 (0.30)	0.664	0.23 (0.30)	0.443
mmp9	-0.14 (0.19)	0.581	-0.08 (0.23)	0.733	-0.09 (0.23)	0.687	-0.09 (0.23)	0.690
cyto5.6	-0.60 (0.25)	0.014	-0.74 (0.33)	0.025	-0.67 (0.30)	0.025	-0.79 (0.36)	0.028
vim	-0.27 (0.22)	0.233	-0.30 (0.27)	0.269	-0.29 (0.27)	0.272	-0.31 (0.27)	0.252
SMA	0.37 (0.26)	0.155	0.53 (0.35)	0.134	0.42 (0.34)	0.214	0.57 (0.36)	0.111
zeb1	0.55 (0.21)	0.010	0.66 (0.26)	0.012	0.62 (0.25)	0.013	0.73 (0.28)	0.010
zeb2	0.01 (0.26)	0.969	0.03 (0.31)	0.920	-0.02 (0.30)	0.958	-0.02 (0.34)	0.948
S100A	0.14 (0.26)	0.605	0.15 (0.36)	0.685	0.16 (0.35)	0.646	0.21 (0.38)	0.587

Tabella 4.3: Stima dei coefficienti e relativi *standard errors* ottenuti con i due metodi: *PLE* e *ARE*. (Modello completo)

Si procede quindi con la regressione all'indietro eliminando, dal modello completo, di volta in volta i coefficienti che risultano non significativi.

Si riporta quindi in Tabella 4.4, il modello ridotto così ottenuto; i soli parametri significativi risultano essere `site`, `cyto5.6` e `zeb1`, come nell'analisi del capitolo precedente.

Modello ridotto	**PLE**		**ARE** τ=0.95					
			Pesi lineari		**Pesi esponenziali**		**Pesi quadratici**	
variabile	***coef (s.e.)***	***p-value***	***coef (s.e.)***	***p-value***	***coef (s.e.)***	***p-value***	***Coef (s.e.)***	***p-value***
`Site Pleural`	-0.61 (0.31)	0.052	-0.74 (0.37)	0.045	-0.70 (0.36)	0.050	-0.85 (0.38)	0.026
`cyto5.6`	-0.42 (0.20)	0.034	-0.55 (0.22)	0.010	-0.52 (0.21)	0.014	-0.64 (0.24)	0.008
`zeb1`	0.36 (0.15)	0.016	0.40 (0.19)	0.030	0.38 (0.19)	0.040	0.42 (0.19)	0.028

Tabella 4.4: Stima dei coefficienti e relativi *standard errors* ottenuti con i due metodi: *PLE* e *ARE*. (Modello ridotto)

Anche nel modello di Cox a rischi proporzionali stimato nel § 3.4 si sono ottenuti risultati simili procedendo con l'eliminazione delle variabili non significative.
In Tabella 4.4 è possibile confrontare i due modelli ridotti per *PLE* e *ARE*.
Si nota quindi che, per la stima *ARE*, il *p-value* della variabile `sitePleural` è molto più affidabile del modello equivalente *PLE*. Infatti in *PLE*, la stima di `sitePleural` è -0.61 che si riconduce a un *hazard ratio* pari a exp(-0.61) = 0.54 (p = 0.052) mentre per *ARE*, utilizzando i pesi quadratici, la stima robusta che si ottiene è -0.85 che si riconduce a un *hazard ratio* pari a exp(-0.85) = 0.43 con un *p-value* (p=0.026) particolarmente significativo che porta a rifiutare l'ipotesi di nullità del coefficiente al 5%.

Per evidenziare i potenziali *outliers*, è possibile ottenere il grafico dei pesi su scala logaritmica, calcolati come

$$-K\log(w_i) = \lambda_0(t_i)\exp(x_i^T\beta),$$

sul numero dei pazienti. E' ragionevole quindi utilizzare il modello esponenziale per ottenere il grafico dei pesi robusti *ARE* (log-trasformati) sul numero di pazienti (Figura 4.2).
Scelto il 95% come valore specifico per il quantile τ della distribuzione empirica, attraverso l'implementazione di *R*, si ricava il valore di troncamento *K* pari a 2.34.
Nella Figura 4.2 i punti di colore nero rappresentano i pesi calcolati per i pazienti deceduti mentre i punti bianchi rappresentano i dati censurati. Per costruzione, avendo

scelto $\tau=0.95$, ci si aspetta che all'incirca il 95% dei pesi si trovi al di sotto di tale riferimento.

Si può notare che solo il 2,3% delle osservazioni è posizionato al di sopra della linea orizzontale che, per l'appunto, rappresenta *K*: questi tre casi in cui i pesi robusti *ARE* (log-trasformati) si scostano dal gruppo sono riconducibili ai pesi calcolati per i pazienti 51, 40 e 46. Vi sono invece due casi (corrispondenti ai pazienti 26 e 70) in cui il valore dei pesi robusti *ARE* (log-trasformati) coincidono con *K*.

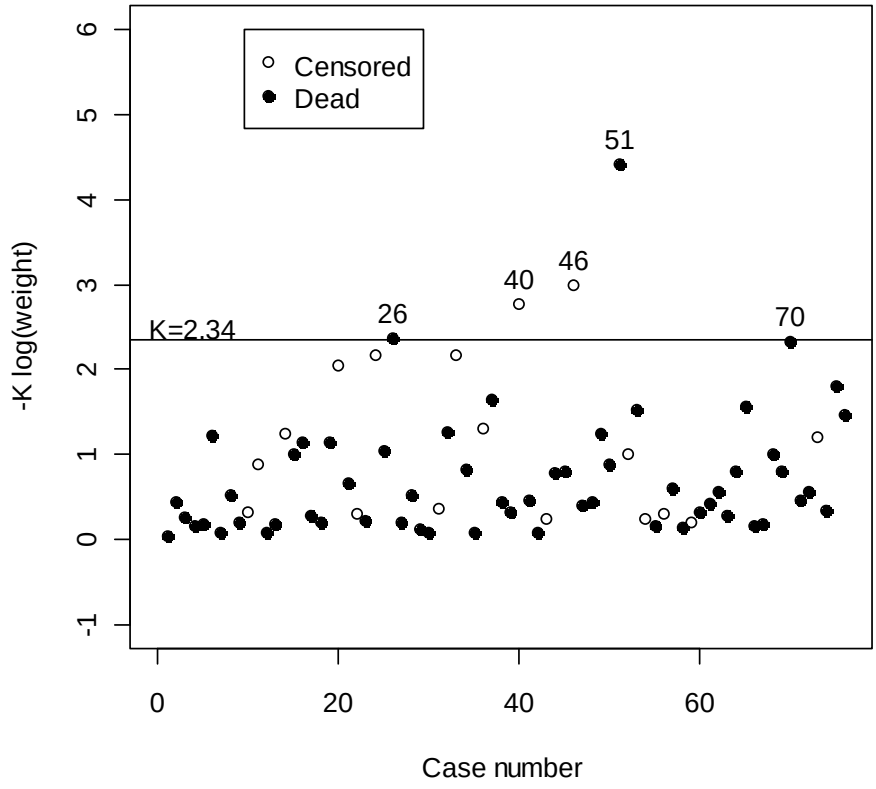

Figura 4.2: Grafico dei pesi esponenziali *ARE* (log-trasformati) sul numero di pazienti.

Per effettuare un'ulteriore analisi grafica del modello ridotto *ARE*, a confronto con il modello *PLE*, si utilizzano i pesi quadratici.

I risultati numerici della Tabella 4.4 sono quindi supportati da cinque grafici: il primo (Figura 4.3) permette di confrontare come i dati sono spiegati dal modello stimato dei rischi proporzionali - non robusto (linea di colore nero) e dal metodo robusto (linea di colore verde).

Gli altri quattro grafici in Figura 4.4, invece, rappresentano le differenze nei quattro strati, definiti dai quantili del predittore lineare.

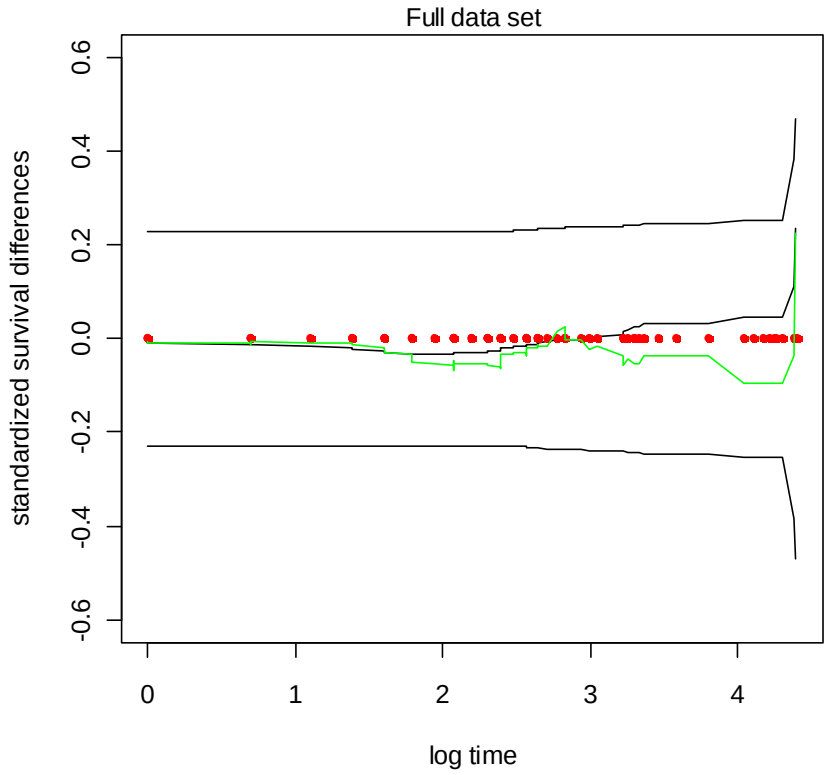

Figura 4.3: Grafico della stima *ARE* con stima dei pesi quadratici.

Il grafico (Figura 4.3) della stima *ARE*, con stima dei pesi quadratici, evidenzia che il metodo robusto (linea di colore verde) si adatta maggiormente ai dati (punti in rosso) rispetto al modello dei rischi proporzionali - non robusto (linea di colore nero).
Infatti, dai quattro grafici della Figura 4.4 si può notare il buon adattamento della curva di sopravvivenza stimata tramite il modello *ARE* (curva verde) mentre, l'altra stimata tramite lo stimatore di Kaplan-Meier (curva nera) tende ai limiti, indice di inadeguatezza.

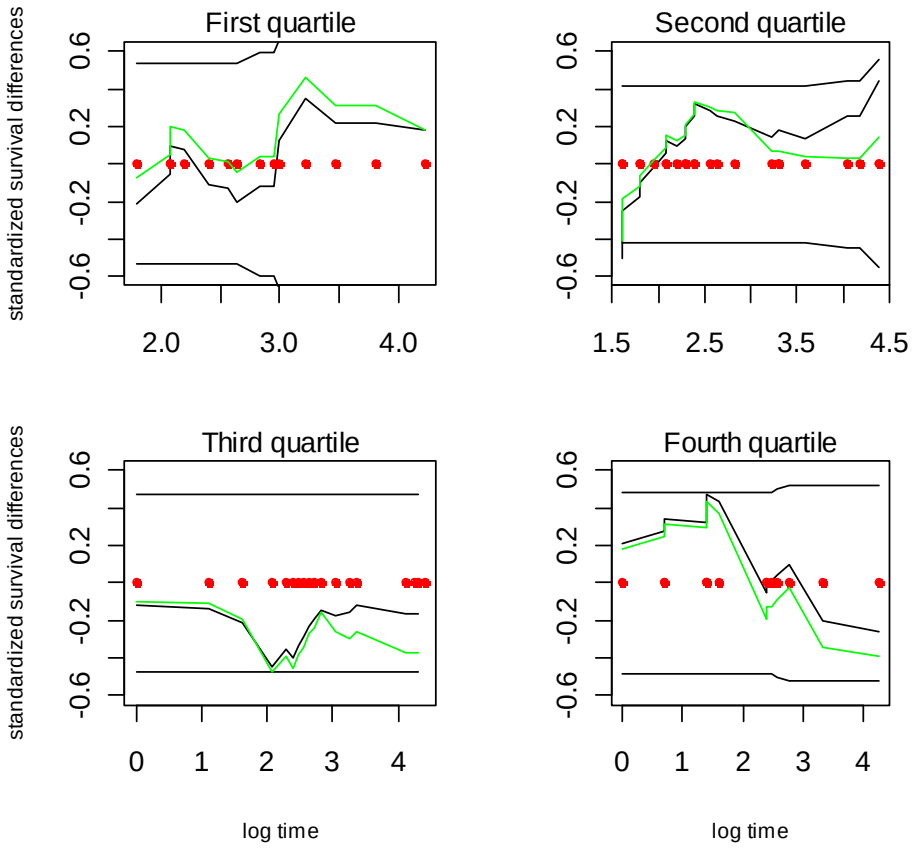

Figura 4.4: Differenza standardizzata di due funzioni di sopravvivenza stimata, una tramite lo stimatore di Kaplan-Meier e l'altra tramite il modello *ARE*.

4.5 Conclusioni

In questo capitolo è stata presentata la tecnica di regressione robusta *adaptive robust estimator* che, per la semplicità di analisi e comprensione, può essere considerata una valida alternativa all'analisi della sopravvivenza vista nel Capitolo 3.

Nel paragrafo 4.4, è stata quindi applicata la teoria di Bednarski al *dataset* ottenendo risultati più che soddisfacenti dal confronto con il modello a rischi proporzionali di Cox.

CONCLUSIONI

Obiettivo principale di questa tesi è analizzare la sopravvivenza dei pazienti a cui è stato diagnosticato il mesotelioma maligno, grave forma di cancro correlata all'esposizione alle fibre aerodisperse dell'amianto.

Dopo una presentazione generale del tumore e delle caratteristiche che lo contraddistinguono (Capitolo 1), è stata effettuata un'analisi esplorativa di tutte le variabili osservate per i pazienti (Capitolo 2).

Si è quindi approfondito lo studio per esaminare quali fattori hanno maggior influenza sulla variabile dipendente `survival` (Capitolo 3), effettuando l'analisi dei tempi di sopravvivenza, con dati censurati.

Si è ritenuto opportuno ricorre al modello semi-parametrico di *Cox* a rischi proporzionali. L'analisi ha complessivamente portato a stimare i tempi di sopravvivenza con un discreto modello che vede interessate nella regressione le tre variabili esplicative `site`, `cyto5.6` e `zeb1`. Si ricorda che, la variabile `site` è determinata dal distretto corporeo nel quale ha avuto origine il MM (pleurico o peritoneale) mentre le altre due variabili rappresentano il *citoscheletro* (`cyto5.6`), proteina che svolge un ruolo indispensabile nella migrazione mesenchimale (§ 1.3.1), e il fattore di trascrizione *zinc finger* (`zeb1`).

L'analisi dei residui di tale modello ha evidenziato però alcuni *outlier*s. Poichè in letteratura il modello di Cox è considerato limitato e talvolta inadeguato nel caso di dati anomali, si è ritenuto opportuno ricorrere al modello con i "*Robustly Proportional Hazards*". Questo approccio robusto rispetto agli *outliers* o a eventuali specificazioni scorrette (Heritier *et al.*, 2009), trattato nel Capitolo 4, ha portato risultati più affidabili rispetto al modello di Cox.

Entrambi i modelli, portano a constatare che le tre variabili esplicative `site`, `cyto5.6` e `zeb1` risultano di particolare interesse per lo studio del dataset ma, confrontando le stime, per il modello robusto si ottengono *p-values* più stabili rispetto il modello equivalente a rischi proporzionali.

Ciò è confermato anche da opportune analisi grafiche (§ 4.4) in cui è evidente che il metodo robusto si adatta meglio ai dati in esame rispetto al modello dei rischi proporzionali - non robusto. Inoltre, il grafico dei pesi esponenziali, calcolati con metodo robusto, in scala logaritmica sul numero dei pazienti non ha evidenziato

potenziali *outliers*, dimostrando che la probabile presenza di alcuni casi anomali riscontrati per il modello di Cox, non è del tutto fondata.

Si può quindi concludere che, per l'analisi di *datasets* provenienti da studi medici nella quale si possono riscontrare pazienti particolari che si distinguono da altri per caratteristiche o per evoluzione della patologia in modo differente e pertanto non eliminabili, è preferibile utilizzare il modello di regressione stimato per l'appunto con i "*Robustly Proportional Hazards*".

APPENDICE A

Dataset utilizzato

Tabella A.1: Dataset utilizzato per l'analisi dei pazienti.

	sesso	site	subtype	survival	status	eta	ecad	ncad	bcad	mmp2	mmp9	cyto5.6	vim	SMA	zeb1	zeb2	S100A
1	M	Peritoneal	Sarcomatous	0	deceduto	76	0	3	1	1	3	0	3	2	3	1	2
2	M	Pleural	Mixed	11	deceduto	72	1	2	2	1	3	1	2	1	2	1	2
3	F	Pleural	Epithelioid	8	deceduto	49	3	2	2	1	1	1	3	0	2	0	1
4	F	Peritoneal	Epithelioid	2	deceduto	58	3	2	2	1	1	0	1	0	1	0	1
5	M	Pleural	Sarcomatous	3	deceduto	72	0	1	1	0	0	0	3	2	2	0	3
6	F	Pleural	Epithelioid	17	deceduto	71	1	2	2	1	1	0	2	0	2	0	1
7	F	Peritoneal	Mixed	0	deceduto	78	0	0	0	0	2	0	3	1	3	1	1
8	F	Pleural	Mixed	11	deceduto	67	1	3	2	0	0	0	3	2	1	1	2
9	M	Pleural	Epithelioid	6	deceduto	69	1	2	2	0	2	1	2	1	2	1	2
10	M	Pleural	Epithelioid	14	vivo	65	2	2	2	0	2	3	0	0	3	1	1
11	F	Pleural	Epithelioid	68	vivo	70	0	0	0	0	1	1	0	0	1	0	1
12	F	Peritoneal	Epithelioid	1	deceduto	51	0	0	0	1	1	1	1	1	1	0	2
13	M	Pleural	Sarcomatous	5	deceduto	55	0	2	1	1	2	0	3	1	1	1	2
14	M	Pleural	Sarcomatous	12	vivo	71	0	3	2	2	3	0	3	0	3	0	1
15	M	Pleural	Mixed	15	deceduto	70	0	1	2	1	1	1	3	1	3	1	0
16	F	Pleural	Mixed	11	deceduto	76	0	3	1	2	2	0	3	2	3	1	1
17	M	Pleural	Sarcomatous	4	deceduto	68	0	2	1	1	1	0	3	2	3	2	2
18	M	Pleural	Sarcomatous	2	deceduto	69	0	2	1	0	0	0	3	1	3	1	1
19	M	Pleural	Mixed	27	deceduto	81	3	3	3	1	1	1	3	0	2	0	2
20	M	Peritoneal	Epithelioid	61	vivo	71	1	0	0	0	0	0	0	0	0	0	0
21	F	Pleural	Mixed	25	deceduto	47	3	3	3	0	2	1	2	0	1	0	2
22	M	Pleural	Epithelioid	45	vivo	69	3	1	3	0	0	3	1	0	1	0	0
23	M	Pleural	Epithelioid	8	deceduto	61	3	2	3	0	0	0	1	0	0	0	1
24	M	Pleural	Sarcomatous	74	vivo	76	0	2	1	0	0	0	3	0	2	0	3
25	F	Pleural	Sarcomatous	14	deceduto	78	0	3	0	0	0	0	2	1	2	2	2
26	M	Peritoneal	Epithelioid	16	deceduto	83	1	0	1	0	2	0	2	0	2	1	2
27	M	Pleural	Epithelioid	8	deceduto	79	1	2	3	0	0	1	1	0	1	0	0
28	F	Peritoneal	Epithelioid	10	deceduto	74	1	3	3	0	0	1	1	0	1	0	1
29	F	Pleural	Epithelioid	5	deceduto	81	3	2	3	0	1	0	1	0	0	0	1
30	F	Peritoneal	Epithelioid	0	deceduto	47	0	2	3	3	1	0	2	0	2	0	1
31	F	Pleural	Epithelioid	13	vivo	61	2	0	3	0	3	1	0	0	1	0	0
32	M	Peritoneal	Epithelioid	21	deceduto	66	0	0	0	0	1	1	1	1	1	0	1
33	F	Pleural	Sarcomatous	70	vivo	83	0	2	1	3	2	0	3	3	2	2	2
34	M	Pleural	Epithelioid	11	deceduto	64	0	2	0	0	2	0	2	1	2	1	2
35	F	Pleural	Epithelioid	0	deceduto	81	0	1	0	1	1	0	2	0	3	0	0
36	M	Pleural	Epithelioid	65	vivo	68	3	2	2	1	1	1	3	0	2	0	1
37	M	Peritoneal	Mixed	29	deceduto	82	1	0	1	0	2	1	3	1	1	1	3
38	F	Pleural	Epithelioid	13	deceduto	76	3	2	3	0	1	0	1	0	0	0	1
39	M	Peritoneal	Epithelioid	5	deceduto	83	0	1	1	0	1	2	3	2	3	0	1
40	M	Pleural	Mixed	28	vivo	66	1	3	3	0	1	0	2	0	3	0	2

	sesso	site	subtype	survival	status	eta	ecad	ncad	bcad	mmp2	mmp9	cyto5.6	vim	SMA	zeb1	zeb2	S100A
41	M	Peritoneal	Sarcomatous	2	deceduto	81	0	0	0	2	2	0	3	1	3	1	1
42	F	Pleural	Epithelioid	6	deceduto	59	1	2	1	1	1	3	1	0	2	1	2
43	M	Pleural	Epithelioid	9	vivo	64	1	2	3	0	0	1	1	0	1	0	0
44	M	Pleural	Epithelioid	17	deceduto	71	0	0	0	1	1	1	1	0	2	1	1
45	M	Peritoneal	Epithelioid	13	deceduto	64	1	1	1	2	0	1	1	1	1	0	0
46	F	Peritoneal	Epithelioid	71	vivo	74	2	0	1	0	1	0	2	0	1	1	1
47	M	Pleural	Epithelioid	9	deceduto	68	3	3	3	0	0	0	0	0	1	0	0
48	M	Pleural	Mixed	10	deceduto	75	2	3	3	3	2	0	2	0	1	0	2
49	M	Pleural	Sarcomatous	11	deceduto	75	0	1	0	0	0	0	3	2	3	1	3
50	M	Pleural	Mixed	12	deceduto	69	0	2	1	1	1	0	3	2	2	2	3
51	F	Pleural	Epithelioid	82	deceduto	63	0	3	2	0	1	0	3	0	2	0	1
52	M	Peritoneal	Epithelioid	27	vivo	57	1	1	2	1	2	1	2	0	0	2	2
53	M	Peritoneal	Epithelioid	8	deceduto	75	0	1	0	0	1	0	3	0	3	0	0
54	M	Pleural	Epithelioid	11	vivo	53	2	2	2	1	0	2	2	0	2	0	1
55	M	Pleural	Epithelioid	8	deceduto	74	3	0	2	0	0	1	2	0	0	1	2
56	M	Pleural	Epithelioid	32	vivo	71	3	1	3	0	0	3	1	0	1	0	0
57	F	Pleural	Epithelioid	20	deceduto	68	1	1	0	0	1	1	2	0	1	0	1
58	F	Peritoneal	Sarcomatous	0	deceduto	77	0	1	0	1	2	0	3	1	2	0	3
59	M	Pleural	Epithelioid	19	vivo	64	1	1	1	0	0	3	1	0	1	1	0
60	M	Pleural	Sarcomatous	4	deceduto	72	1	3	2	0	1	0	3	2	3	2	2
61	M	Pleural	Epithelioid	10	deceduto	79	0	0	0	1	0	1	1	0	2	2	1
62	M	Pleural	Sarcomatous	8	deceduto	61	0	2	1	0	1	0	3	1	2	1	2
63	F	Pleural	Epithelioid	7	deceduto	83	3	2	2	1	1	0	1	0	1	0	1
64	F	Peritoneal	Epithelioid	14	deceduto	55	1	0	1	0	0	2	2	1	2	1	1
65	M	Pleural	Mixed	26	deceduto	62	0	2	2	2	1	1	3	2	3	1	2
66	M	Pleural	Sarcomatous	1	deceduto	66	0	2	0	0	0	0	3	1	3	1	2
67	M	Pleural	Epithelioid	6	deceduto	67	3	2	3	0	0	0	1	0	0	0	1
68	M	Pleural	Epithelioid	57	deceduto	68	0	1	2	1	0	0	2	0	0	2	1
69	F	Pleural	Epithelioid	25	deceduto	58	2	3	3	0	1	0	3	0	0	0	2
70	M	Pleural	Epithelioid	81	deceduto	66	1	1	1	0	1	0	2	0	1	0	1
71	M	Pleural	Epithelioid	5	deceduto	77	0	1	2	1	0	0	2	0	3	1	1
72	M	Pleural	Mixed	17	deceduto	72	0	1	2	0	2	1	1	0	1	0	1
73	M	Pleural	Mixed	36	vivo	77	1	1	1	0	0	1	2	1	2	0	2
74	M	Pleural	Epithelioid	4	deceduto	68	0	1	2	1	0	0	2	0	3	1	1
75	M	Pleural	Epithelioid	80	deceduto	73	0	2	1	0	2	0	1	0	1	0	1
76	M	Pleural	Sarcomatous	13	deceduto	75	0	3	2	1	1	0	2	0	3	1	1

APPENDICE B

Comandi e output R per lo studio della sopravvivenza

Per una maggiore comprensione del Capitolo 3, vengono riportati di seguito i comandi più significativi, e i relativi output di *R*, che sono stati utilizzati per lo studio della sopravvivenza.

Dal *dataset* iniziale (Tabella A.1 – Appendice A), viene codificata la variabile dicotomica `status`. Si assume:

- `status = 1`, ovvero i pazienti deceduti (dati non censurati);
- `status = 0`, ovvero i pazienti vivi al 31/12/10 (dati censurati).

Quindi viene richiamata in *R* la "`library(survival)`" :

```
> Surv(survival,status)
 [1]  0  11   8   2   3  17   0  11   6  14+ 68+  1   5  12+ 15  11   4   2
[19] 27  61+ 25  45+  8  74+ 14  16   8  10   5   0  13+ 21  70+ 11   0  65+
[37] 29  13   5  28+  2   6   9+ 17  13  71+  9  10  11  12  82  27+  8  11+
[55]  8  32+ 20   0  19+  4  10   8   7  14  26   1   6  57  25  81   5  17
[73] 36+  4  80  13

> summary(as.factor(status))
 0  1
17  9
```

Per calcolare la stima di Kaplan - Meier (per i 2 gruppi di `site`):

```
> KM=survfit(Surv(survival,status)~site)

> KM
Call: survfit(formula = Surv(survival, status) ~ site)

                records n.max n.start events median 0.95LCL 0.95UCL
site=Peritoneal      18    18      18     15      9       2      29
site=Pleural         58    58      58     44     13      11      25
```

Per ottenere il grafico delle stime di Kaplan Meier:

```
> plot(KM, xlab ="t(mesi)", ylab= "S(t)",col=1:2,lty=2:3)
> legend(60,.8,c("peritoneal","pleural"),col=2:1,lty=2:3)
```

Log-rank test per confrontare le curve di sopravvivenza:

```
> survdiff(Surv(survival,status)~site)
Call:
survdiff(formula = Surv(survival, status) ~ site)
                 N Observed Expected (O-E)^2/E (O-E)^2/V
site=Peritoneal 18       15     11.1      1.33      1.75
site=Pleural    58       44     47.9      0.31      1.75
 Chisq= 1.7  on 1 degrees of freedom, p= 0.186
```

Per calcolare la stima di Kaplan - Meier (per i 3 sottogruppi di `subtype`):

```
> KM=survfit(Surv(survival,status)~subtype)
> KM
Call: survfit(formula = Surv(survival, status) ~ subtype)

                 records n.max n.start events median 0.95LCL 0.95UCL
subtype=Epithelioid   46    46      46     34   13.0       9      25
subtype=Mixed         14    14      14     12   16.0      11      NA
subtype=Sarcomatous   16    16      16     13    4.5       2      NA
```

Per ottenere il grafico delle stime di Kaplan Meier:

```
> plot(KM, xlab ="t(mesi)", ylab= "S(t)",
  col=c("pink","blue","green"))

> legend(60,.8,c("epithelioid","mixed","sarcomatous"),
  col=c("pink","blue","green"),lty=2:3)
```

Log-rank test per confrontare le curve di sopravvivenza:

```
> survdiff(Surv(survival,status)~subtype)
Call:
survdiff(formula = Surv(survival, status) ~ subtype)

                   N Observed Expected (O-E)^2/E (O-E)^2/V
subtype=Epithelioid 46       34    38.39    0.5019      1.56
subtype=Mixed       14       12    12.97    0.0722      0.10
subtype=Sarcomatous 16       13     7.64    3.7543      4.61

 Chisq= 4.6  on 2 degrees of freedom, p= 0.0988
```

Modello di Cox completo:

```
> summary(mod.cox)

Call:

coxph(formula = Surv(survival, status) ~ site + subtype + sesso + eta
+ ecad + ncad + bcad + mmp2 + mmp9 + cyto5.6 + vim + SMA + zeb1 + zeb2
+ S100A)

  n= 76

                       coef exp(coef)  se(coef)      z Pr(>|z|)
sitePleural        -0.820237  0.440327  0.412645 -1.988  0.04684 *
subtypeMixed       -0.406513  0.665969  0.520625 -0.781  0.43491
subtypeSarcomatous  0.060025  1.061863  0.587084  0.102  0.91856
sessoM             -0.013202  0.986885  0.317660 -0.042  0.96685
eta                -0.023219  0.977048  0.019836 -1.171  0.24178
ecad                0.214963  1.239816  0.196683  1.093  0.27442
ncad                0.009915  1.009964  0.224027  0.044  0.96470
bcad                0.071595  1.074220  0.246028  0.291  0.77105
mmp2                0.074072  1.076885  0.215465  0.344  0.73101
mmp9               -0.127140  0.880611  0.187806 -0.677  0.49842
cyto5.6            -0.607456  0.544735  0.246428 -2.465  0.01370 *
vim                -0.255479  0.774546  0.223756 -1.142  0.25355
SMA                 0.360745  1.434397  0.260549  1.385  0.16619
zeb1                0.553829  1.739903  0.214369  2.584  0.00978 **
zeb2                0.009836  1.009884  0.256579  0.038  0.96942
S100A               0.142586  1.153253  0.262223  0.544  0.58661
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                   exp(coef) exp(-coef) lower .95 upper .95
sitePleural           0.4403     2.2710    0.1961    0.9886
subtypeMixed          0.6660     1.5016    0.2400    1.8476
subtypeSarcomatous    1.0619     0.9417    0.3360    3.3558
sessoM                0.9869     1.0133    0.5295    1.8393
eta                   0.9770     1.0235    0.9398    1.0158
ecad                  1.2398     0.8066    0.8432    1.8229
ncad                  1.0100     0.9901    0.6510    1.5667
bcad                  1.0742     0.9309    0.6632    1.7399
mmp2                  1.0769     0.9286    0.7059    1.6428
mmp9                  0.8806     1.1356    0.6094    1.2725
cyto5.6               0.5447     1.8358    0.3361    0.8830
vim                   0.7745     1.2911    0.4996    1.2009
SMA                   1.4344     0.6972    0.8608    2.3903
zeb1                  1.7399     0.5747    1.1430    2.6485
zeb2                  1.0099     0.9902    0.6108    1.6698
S100A                 1.1533     0.8671    0.6898    1.9281

Rsquare= 0.239   (max possible= 0.996 )
Likelihood ratio test= 20.79  on 16 df,   p=0.1866
Wald test            = 18.67  on 16 df,   p=0.2863
Score (logrank) test = 19.94  on 16 df,   p=0.2229
```

Modello di Cox ridotto:

```
> summary(mod.coxR)

Call:
coxph(formula = Surv(survival, status) ~ site + cyto5.6 + zeb1)
  n= 76
              coef exp(coef) se(coef)      z Pr(>|z|)
```

```
sitePleural -0.6182    0.5389   0.3163 -1.955   0.0506 .
cyto5.6     -0.4284    0.6516   0.1988 -2.155   0.0312 *
zeb1         0.3680    1.4449   0.1509  2.439   0.0147 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
            exp(coef) exp(-coef) lower .95 upper .95
sitePleural    0.5389     1.8556    0.2899     1.002
cyto5.6        0.6516     1.5348    0.4413     0.962
zeb1           1.4449     0.6921    1.0750     1.942

Rsquare= 0.165   (max possible= 0.996 )
Likelihood ratio test= 13.71  on 3 df,   p=0.003328
Wald test            = 12.75  on 3 df,   p=0.005218
Score (logrank) test = 12.89  on 3 df,   p=0.004883
```

Confronto tra i due modelli:

```
> anova(mod.coxR, mod.cox, test="Chisq")
Analysis of Deviance Table
Model 1: Surv(survival, status) ~ site + cyto5.6 + zeb1
Model 2: Surv(survival, status) ~ site + subtype + sesso + eta + ecad
+ncad + bcad + mmp2 + mmp9 + cyto5.6 + vim + SMA + zeb1 + zeb2 + S100A
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        73     408.78
2        60     401.69 13     7.08      0.90
```

Modello di Cox stratificato per la variabile *site*

```
> summary(cox.strat)
Call:
coxph(formula = Surv(survival, status) ~ strata(site) + cyto5.6 +
zeb1)
  n= 76
          coef exp(coef) se(coef)      z Pr(>|z|)
cyto5.6 -0.4257    0.6533   0.1978 -2.152   0.0314 *
zeb1     0.3114    1.3653   0.1500  2.076   0.0379 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
cyto5.6    0.6533     1.5306    0.4434    0.9628
zeb1       1.3653     0.7324    1.0175    1.8320

Rsquare= 0.128   (max possible= 0.991 )
Likelihood ratio test= 10.41  on 2 df,  p=0.005477
Wald test            = 9.55  on 2 df,   p=0.008417
Score (logrank) test = 9.78  on 2 df,   p=0.007507
```

Per visualizzare campioni per tipologia:

```
> res = survfit(Surv(survival, status) ~ as.factor(site))
> res$strata
 as.factor(site)=Peritoneal    as.factor(site)=Pleural
                         14                         34
```

Grafico curve di sopravvivenza per strato:

```
> strato=c(rep(1,14),rep(2,34))
> plot(res$time, -log(res$surv), type= 'n', col=c("red","blue"))
> legend(1,2,c("Peritoneal","Pleural"), col=c("red","blue"),lty=1:1)
> lines(res$time[strato==1],-log(res$surv[strato==1]),type='s',
  col=2)
> lines(res$time[strato==2],-log(res$surv[strato==2]),type='s',
  col=4)
```

Test rischio proporzionale modello di Cox ridotto:

```
> cox.zph(mod.coxR)
               rho chisq      p
sitePleural  0.224  3.43 0.0642
cyto5.6      0.206  2.28 0.1314
zeb1        -0.186  2.70 0.1003
GLOBAL          NA  7.17 0.0666
```

Residui modello di Cox ridotto:

1. *residui di martingala:*

```
> res.m=residuals(mod.coxR,type='mart')
> res.cs=status-res.m
> s.res=survfit(Surv(res.cs,status))
> plot(s.res$time,-log(s.res$surv), type='s')
```

2. *residui di devianza:*

```
> resi.d=residuals(mod.coxR,type='dev')
> plot(res.d, main="residui di devianza")
> qqnorm(resi.d)
> qqline(resi.d)
```

3. *residui parziali di Schoenfeld:*

```
> cox.res.S=cox.zph(mod.coxR)
```

```
> cox.res.S
               rho chisq      p
sitePleural  0.224  3.43 0.0642
cyto5.6      0.206  2.28 0.1314
zeb1        -0.186  2.70 0.1003
GLOBAL          NA  7.17 0.0666
> plot(cox.res.S)
```

4. residui a score:

```
> r.score = residuals(mod.coxR, type="score")
> plot(r.score[,1], main="res. sitePleural")
> abline(0,0, lty=5)
> identify(r.score[,1])
> plot(r.score[,2], main="res. cyto5.6")
> abline(0,0, lty=5)
> plot(r.score[,3], main="res. zeb1")
> abline(0,0, lty=5)
> r.score
  sitePleural       cyto5.6         zeb1
1  -0.59236723 -0.3083869937  0.833811467
2   0.11897047  0.2424918916  0.120870045
     …                …           …
75 -0.36742562  0.6010869692  0.591706208
76 -0.05018867  0.0581276394 -0.215486934
```

5. residui "Beta":

```
> resB= residuals(mod.coxR, type="dfbetas")
> plot(resB[,1], main="res. sitePleural")
> abline(0,0, lty=5)
> identify(resB[,1])
> plot(resB[,2], main="res. cyto5.6" )
> abline(0,0, lty=5)
> identify(resB[,2])
> plot(resB[,3], main="res. zeb1")
```

```
> abline(0,0, lty=5)
> identify(resB[,3])
> resB
            [,1]         [,2]         [,3]
1  -0.2253042699 -0.058180574  0.176030571
2   0.0336217836  0.050356116  0.010448404
             (…)          (…)          (…)
75 -0.1388681335  0.122145851  0.130489917
76 -0.0061768664  0.009062588 -0.027194346

> qqnorm(resB[,1], main="Q-Q plot sitePleural")
> qqline(resB[,1])
> qqnorm(resB[,2], main="Q-Q plot cyto5.6")
> qqline(resB[,2])
> qqnorm(resB[,3], main="Q-Q plot zeb1")
> qqline(resB[,3])
```

APPENDICE C

Comandi e output R per l'analisi robusta

Nell'ambiente statistico *R* è disponibile la libreria `coxrobust` che permette di effettuare l'analisi robusta del modello di Cox.

Vengono riportati di seguito i comandi più significativi, e i relativi output, utilizzati per l'analisi robusta affrontata nel Capitolo 4.

```
> library(coxrobust)
> s = Surv(survival,status)
> s
  [1] 0  11   8   2   3  17   0  11   6  14+ 68+  1   5  12+ 15
[16] 11   4   2  27  61+ 25  45+  8  74+ 14  16   8  10   5   0
[31] 13+ 21  70+ 11   0  65+ 29  13   5  28+  2   6   9+ 17  13
[46] 71+  9  10  11  12  82  27+  8  11+  8  32+ 20   0  19+  4
[61] 10   8   7  14  26   1   6  57  25  81   5  17  36+  4  80
[76] 13
```

Modello completo di regressione stimato con i "*Robustly Proportional Hazards*", nominato successivamente come *ARE*, stima quadratica dei pesi, τ =0.95:

```
Call:
coxr(formula = s ~ eta + sesso + site + subtype + ecad + ncad + bcad +
mmp2 + mmp9 + cyto5.6 + vim + SMA + zeb1 + zeb2 + S100A, data = IHC,
trunc = 0.95, singular.ok = TRUE, model = FALSE)

Partial likelihood estimator
                       coef exp(coef) se(coef)       p
eta                -0.02373     0.977   0.0198 0.23186
sessoM             -0.02088     0.979   0.3181 0.94767
sitePleural        -0.83817     0.433   0.4175 0.04470
subtypeMixed       -0.40557     0.667   0.5191 0.43466
subtypeSarcomatous  0.06336     1.065   0.5832 0.91348
ecad                0.22375     1.251   0.1972 0.25641
ncad                0.02553     1.026   0.2265 0.91028
bcad                0.07232     1.075   0.2468 0.76951
mmp2                0.06804     1.070   0.2161 0.75285
mmp9               -0.10476     0.901   0.1899 0.58110
cyto5.6            -0.60170     0.548   0.2460 0.01445
vim                -0.26763     0.765   0.2242 0.23267
SMA                 0.37178     1.450   0.2612 0.15460
zeb1                0.55282     1.738   0.2145 0.00995
zeb2                0.00995     1.010   0.2573 0.96914
S100A               0.13577     1.145   0.2628 0.60547

Wald test=18.5 on 16 df, p=0.297

Robust estimator
                     coef exp(coef) se(coef)      p
eta               -0.0151     0.985   0.0294 0.6072
```

```
sessoM             -0.4993     0.607   0.3862 0.1961
sitePleural        -0.9526     0.386   0.6122 0.1197
subtypeMixed       -0.7314     0.481   0.7329 0.3183
subtypeSarcomatous  0.3175     1.374   0.7133 0.6562
ecad                0.2710     1.311   0.2540 0.2859
ncad               -0.1783     0.837   0.2917 0.5411
bcad                0.2775     1.320   0.3246 0.3925
mmp2                0.2327     1.262   0.3036 0.4435
mmp9               -0.0905     0.913   0.2273 0.6904
cyto5.6            -0.7951     0.452   0.3613 0.0278
vim                -0.3060     0.736   0.2674 0.2524
SMA                 0.5755     1.778   0.3607 0.1106
zeb1                0.7350     2.086   0.2842 0.0097
zeb2               -0.0222     0.978   0.3400 0.9479
S100A               0.2048     1.227   0.3774 0.5874

Extended Wald test=41.7 on 16 df, p=0.000438
```

Modello completo *ARE*, stima lineare dei pesi, τ =0.95:

```
Call:
coxr(formula = s ~ eta + sesso + site + subtype + ecad + ncad +
bcad + mmp2 + mmp9 + cyto5.6 + vim + SMA + zeb1 + zeb2 +      S100A,
data = IHC, trunc = 0.95, f.weight = "linear", singular.ok = TRUE,
model = FALSE)


Partial likelihood estimator
                      coef exp(coef) se(coef)       p
eta                -0.02373     0.977   0.0198 0.23186
sessoM             -0.02088     0.979   0.3181 0.94767
sitePleural        -0.83817     0.433   0.4175 0.04470
subtypeMixed       -0.40557     0.667   0.5191 0.43466
subtypeSarcomatous  0.06336     1.065   0.5832 0.91348
ecad                0.22375     1.251   0.1972 0.25641
ncad                0.02553     1.026   0.2265 0.91028
bcad                0.07232     1.075   0.2468 0.76951
mmp2                0.06804     1.070   0.2161 0.75285
mmp9               -0.10476     0.901   0.1899 0.58110
cyto5.6            -0.60170     0.548   0.2460 0.01445
vim                -0.26763     0.765   0.2242 0.23267
SMA                 0.37178     1.450   0.2612 0.15460
zeb1                0.55282     1.738   0.2145 0.00995
zeb2                0.00995     1.010   0.2573 0.96914
S100A               0.13577     1.145   0.2628 0.60547

Wald test=18.5 on 16 df, p=0.297

Robust estimator
                     coef exp(coef) se(coef)      p
eta                -0.0157     0.984   0.0283 0.5781
sessoM             -0.4190     0.658   0.3696 0.2569
sitePleural        -0.9613     0.382   0.5929 0.1049
subtypeMixed       -0.6180     0.539   0.6833 0.3658
subtypeSarcomatous  0.2545     1.290   0.7075 0.7191
ecad                0.2538     1.289   0.2368 0.2838
ncad               -0.0924     0.912   0.2882 0.7484
bcad                0.2364     1.267   0.3132 0.4502
mmp2                0.1906     1.210   0.2990 0.5238
mmp9               -0.0772     0.926   0.2265 0.7333
cyto5.6            -0.7384     0.478   0.3302 0.0253
```

```
vim                -0.3015     0.740   0.2731 0.2697
SMA                 0.5313     1.701   0.3546 0.1341
zeb1                0.6625     1.940   0.2627 0.0117
zeb2                0.0316     1.032   0.3158 0.9203
S100A               0.1459     1.157   0.3639 0.6885

Extended Wald test=34.0 on 16 df, p=0.00551
```

Modello completo *ARE*, stima esponenziale dei pesi, τ =0.95:

```
Call:
coxr(formula = s ~ eta + sesso + site + subtype + ecad + ncad +
bcad + mmp2 + mmp9 + cyto5.6 + vim + SMA + zeb1 + zeb2 +      S100A,
data = IHC, trunc = 0.95, f.weight = "exp", singular.ok = TRUE,
model = FALSE)

Partial likelihood estimator
                       coef exp(coef) se(coef)       p
eta                -0.02373     0.977   0.0198 0.23186
sessoM             -0.02088     0.979   0.3181 0.94767
sitePleural        -0.83817     0.433   0.4175 0.04470
subtypeMixed       -0.40557     0.667   0.5191 0.43466
subtypeSarcomatous  0.06336     1.065   0.5832 0.91348
ecad                0.22375     1.251   0.1972 0.25641
ncad                0.02553     1.026   0.2265 0.91028
bcad                0.07232     1.075   0.2468 0.76951
mmp2                0.06804     1.070   0.2161 0.75285
mmp9               -0.10476     0.901   0.1899 0.58110
cyto5.6            -0.60170     0.548   0.2460 0.01445
vim                -0.26763     0.765   0.2242 0.23267
SMA                 0.37178     1.450   0.2612 0.15460
zeb1                0.55282     1.738   0.2145 0.00995
zeb2                0.00995     1.010   0.2573 0.96914
S100A               0.13577     1.145   0.2628 0.60547

Wald test=18.5 on 16 df, p=0.297

Robust estimator
                      coef exp(coef) se(coef)      p
eta                -0.0205     0.980   0.0268 0.4437
sessoM             -0.2163     0.806   0.4768 0.6501
sitePleural        -0.9059     0.404   0.5484 0.0986
subtypeMixed       -0.4804     0.619   0.6713 0.4743
subtypeSarcomatous  0.1853     1.204   0.6779 0.7845
ecad                0.2238     1.251   0.2345 0.3397
ncad               -0.0493     0.952   0.2529 0.8455
bcad                0.1577     1.171   0.3140 0.6154
mmp2                0.1304     1.139   0.2999 0.6636
mmp9               -0.0918     0.912   0.2276 0.6866
cyto5.6            -0.6749     0.509   0.3019 0.0254
vim                -0.2976     0.743   0.2711 0.2723
SMA                 0.4225     1.526   0.3399 0.2139
zeb1                0.6220     1.863   0.2519 0.0135
zeb2               -0.0160     0.984   0.3053 0.9583
S100A               0.1631     1.177   0.3555 0.6464

Extended Wald test=32.1 on 16 df, p=0.00958
```

Modello ARE ridotto, stima quadratica dei pesi, τ =0.95:

```
Call:
coxr(formula = s ~ site + cyto5.6 + zeb1, data = IHC, trunc = 0.95,
singular.ok = TRUE, model = FALSE)

Partial likelihood estimator
              coef exp(coef) se(coef)      p
sitePleural -0.614     0.541    0.316 0.0524
cyto5.6     -0.422     0.656    0.199 0.0338
zeb1         0.363     1.437    0.151 0.0162

Wald test=12.4 on 3 df, p=0.00604

Robust estimator
              coef exp(coef) se(coef)      p
sitePleural -0.849     0.428    0.382 0.0261
cyto5.6     -0.636     0.530    0.241 0.0084
zeb1         0.416     1.516    0.189 0.0280

Extended Wald test=12.6 on 3 df, p=0.00557
```

Modello *ARE* ridotto, stima lineare dei pesi, τ =0.95:

```
Call:
coxr(formula = s ~ site + cyto5.6 + zeb1, data = IHC, trunc = 0.95,
f.weight = "linear", singular.ok = TRUE, model = FALSE)

Partial likelihood estimator
              coef exp(coef) se(coef)      p
sitePleural -0.614     0.541    0.316 0.0524
cyto5.6     -0.422     0.656    0.199 0.0338
zeb1         0.363     1.437    0.151 0.0162

Wald test=12.4 on 3 df, p=0.00604

Robust estimator
              coef exp(coef) se(coef)      p
sitePleural -0.736     0.479    0.367 0.0449
cyto5.6     -0.554     0.575    0.216 0.0103
zeb1         0.404     1.498    0.186 0.0301

Extended Wald test=12.5 on 3 df, p=0.00579
```

Modello *ARE* ridotto, stima esponenziale dei pesi, τ =0.95:

```
Call:
coxr(formula = s ~ site + cyto5.6 + zeb1, data = IHC, trunc = 0.95,
f.weight = "exp", singular.ok = TRUE, model = FALSE)

Partial likelihood estimator
              coef exp(coef) se(coef)      p
sitePleural -0.614     0.541    0.316 0.0524
cyto5.6     -0.422     0.656    0.199 0.0338
zeb1         0.363     1.437    0.151 0.0162

Wald test=12.4 on 3 df, p=0.00604

Robust estimator
```

```
              coef exp(coef) se(coef)      p
sitePleural -0.703     0.495    0.360 0.0507
cyto5.6     -0.516     0.597    0.209 0.0137
zeb1         0.385     1.469    0.188 0.0406

Extended Wald test=11.7 on 3 df, p=0.00861
```

Per il calcolo dei pesi esponenziali secondo la formula

$$-K\log(w_i) = \lambda_0(t_i)\exp(x_i^T\beta),$$

per semplicità, dal dataset iniziale, si crea un ulteriore dataset con le sole variabili d'interesse `survival`, `status` e le tre esplicative risultate significative: `site`, `cyto5.6` e `zeb1`. La variabile dicotomica `site` viene codificata con `peritoneal=0` e `pleural=1`.

Quindi, le tre esplicative vengono prelevate per il calcolo dei pesi:

```
x=IHCrid[,c(1,4,5)]
```

Viene quindi ristimato il modello *ARE* con pesi esponenziali, τ =0.95:

```
> fit.ARE = rcoxph(IHCrid$survival, IHCrid$status, x ,wt.type =
 "exponential", quant=.95)

> fit.ARE$coefficients
          estimate        SE         z p.value
site    -0.7042206 0.3463981 -2.032981 0.04205
cyto5.6 -0.5112682 0.2042217 -2.503496 0.01229
zeb1     0.3837336 0.1560326  2.459316 0.01392
```

Si calcola il valore di troncamento K:

```
> K=quantile(fit.ARE$Lambda.rob*exp(fit.ARE$xbeta.rob),probs=c(.95))
> K
      95%
 2.343335
```

Per ottenere il grafico dei pesi esponenziali *ARE* (log-trasformati) sul numero di pazienti:

```
> nb=seq(1,76,1)
> logwt0=-log(fit.ARE$wt[IHCrid$status==0])*K
> logwt1=-log(fit.ARE$wt[IHCrid$status==1])*K
> n0=nb[IHCrid$status==0]
> n1=nb[IHCrid$status==1]
> plot(-log(fit.ARE$wt)*K, ylab=" -K log(weight)", xlab="Case
  number",xlim=c(0,76),ylim=c(-1,6),type="n")
> points(n0,logwt0,pch=21)
```

```
> points(n1,logwt1,pch=19)
> abline(h=K)
> legend(9.5,6,c("Censored","Dead"),pch=c(21,19))
> text(-2.75,3,"K=2.34",pos=4)
> text(40,2.78,"40",pos=3)
> text(46,2.99,"46",pos=3)
> text(26,2.37,"26",pos=3)
> text(51,4.42,"51",pos=3)
> text(70,2.33,"70",pos=3)
```

Per un confronto grafico di come i dati sono spiegati dal stimato modello ridotto dei rischi proporzionali - non robusto e dal metodo robusto:

```
> plot(coxr)
```

Si ottengono così grafici in Figura 4.3 e 4.4.

BIBLIOGRAFIA

Bianchi, C., Bianchi, T. (2007). Malignant mesothelioma: Global incidence and relationship with asbestos. *Ind.Health*, Vol. 45, pag. 379-387.

Bland, M. (2009). *Statistica Medica*, Apogeo, Milano.

Cappellesso, R. (2009). *"La transazione epiteliale-mesenchimale nel mesotelioma maligno"*. Relatore Prof. A. Fassina, Facoltà di Medicina Chirurgia, Corso di Laurea Specialistica in Medicina e Chirurgia, Università degli Studi di Padova.

Carroll, R.J., Ruppert D. (1988). *Transformation and weighting in regression*. Chapman and Hall, New York.

Desmond, A.F. (1997). Optimal estimating functions, quasi-likelihood and statistical modeling. *Journal of statistical planning and inference*, Vol. 60, pag. 77-104.

Garrett, S.C., Varney, K.M., Weber, D.J., Bresnick, A.R. (2006). S100A4, a mediator of metastasis. *J. Biol. Chem.*, Vol. 281 , pag. 677-680.

Geiger, T.R., Peeper, D.S. (2009). Metastasis mechanisms. *Biochim.Biophys Acta*, Vol. 1796, pag. 293-308.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation. Ph. d Thesis*, University of California, Barkley.

Hampel, F. R. (1971). A general qualitative definition of robustness, *Ann. Math. Stat*, Vol. 42, pag. 1887–1896.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions,* John Wiley & Sons, New York.

Heritier, S., Cantoni, E., Capt, S., Victoria-Feser, M.P. (2009). *Robust Methods in Biostatistics*., Wiley, United Kingdom.

Hirohashi, S. (1998). *Inactivation* of the E-cadherin-mediated cell adhesion system in human cancers, *Am.J.Pathol*, Vol. 153, pag. 333-339.

Huber, P. J. (1964). Robust estimation of a location parameter, *Ann. Math. Stat.*, Vol. 35, pag. 73–101.

Klein, J.P., Moeschberger, M.L. (2003). *Survival analysis: techniques for censored and truncated data*, Springer, New York.

Legge 27 marzo 1992, n. 257 - Norme relative alla cessazione dell'impiego dell'amianto. (pubblicata sul Suppl.Ord. alla Gazzetta Ufficiale n. 87 del 13 aprile 1992) (aggiornata con le modfiche apportate dalla legge 24 aprile 1998, n. 128, della legge 9 dicembre 1998, n. 426, dal decreto-legge 5 giugno 1993, n. 169 e dal decreto- legge 1 ottobre 1996, n. 510).

Marubini E., Valsecchi M.G. (1995). *Analysing survival data from clinical trials and observational studies*, John Wiley & Sons, New York.

Nieman MT, Prudoff RS, Johnson KR, Wheelock MJ. (1999). N-cadherin promotes motility in human breast cancer cells regardless of their E-cadherin expression. *J.Cell Biol*., pag 631-644.

Nieto, M.A. (2002 Mar). The snail superfamily of zinc-finger transcription factors. *Nat.Rev.Mol.Cell Biol*., Vol.3, pag. 155-166.

Pace, L., Salvan, A. (2001). *Introduzione alla Statistica - II Inferenza, verosimiglianza, modelli*, Cedam, Padova.

Peracchi F. (1990). Bounded-influence estimators for the Tobit model, *Journal of Econometrics*, Vol. 44, pag. 107-126.

Piccolo, D. (1998). *Statistica*, Il Mulino, Bologna.

Sawyer, S. (2003). The Greenwood and Exponential Greenwood Confidence Intervals in Survival Analysis. *J. Am. Statist. Assoc*., Vol. 90, pag. 1399-1405.

Schoenfeld, D. (1982). Residuals for the proportional hazards regresssion model. *Biometrika*, Vol. 69, pag. 239-241.

Siegel, S., Castellan Jr, N.J. (1988). *Nonparametric Statistics for The Behavioral Sciences,* McGraw-Hill College, New York.

Storer, B. E., Crowley, J. (1985). A diagnostic for Cox regression and general conditional likelihoods. *J. Am. Statist. Assoc*., Vol. 80, pag. 139–147.

Thiery, J.P. (2002). Epithelial-mesenchymal transitions in tumour progression. *Nat.Rev. Cancer*, Vol. 2, pag. 442-454.

Thiery, J.P, Acloque H, Huang RY, Nieto MA. (2009). Epithelialmesenchymal transitions in development and disease. *Cell*, Vol. 139, pag. 871-890.

Yang J,. Weinberg RA (2008). Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. *Dev.Cell.*, Vol. 14, pag. 818-829.

Link:

Associazione Italiana per la Ricerca sul Cancro, 2011, agg. Giugno 2011, http://www.airc.it/tumori/mesotelioma.asp.

Baccini, M., Mealli F. (2001). *Metodi diagnostici basati sui residui nei modelli per dati di durata*. Università degli Studi di Firenze.
http://www.ds.unifi.it/ricerca/pubblicazioni/altre/.../didattica2001_01.pdf

Grigoletto, F. (2010). *Analisi della sopravvivenza, Statistica Medica per le Scuole Specializzate Università degli Studi di Padova* – Facoltà di Medicina e Chirurgia, Unità didattica n°5.
http://147.162.76.190/didattica/ScuoleSpecializzazione2010/Lezione%20Analisi%20sopravvivenza_220610.pdf

Gruppo Italiano Mesotelioma, *Dati epidemiologici*, 2011, agg. Giugno 2011, http://www.gime.it/clinica03.htm.

Gruppo Italiano Mesotelioma, *Stradiazione e fattori prognostici*, 2011, agg. Giugno 2011, http://www.gime.it/clinica09.htm.

Pelagatti, M. (2000). *L'approccio alla statistica robusta basato sulla funzione d'influenza: appunti per un seminario.* Università degli Studi di Milano-Bicocca.
http://www.statistica.unimib.it/utenti/p_matteo/papers/robusta.pdf

ReNaM*, Il Registro Nazionale dei Mesoteliomi* (DPCM 308/2002), Prefazione di Terracini, B. (Secondo rapporto, 2006).
http://www.ispesl.it/renam/download/Pagine_1_340_secondo_rapp_interno.pdf

Software utilizzato:

Software *R* - Version 2.11.1 (2010-05-31), Copyright (C) 2010 The R Foundation for Statistical Computing, *http://cran.r-project.org/bin/windows/base/old/2.11.0*/